الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

Université Mohamed Khider – Biskra
Faculté des Sciences et de la Technologie
Département: **Génie Electrique**
Ref /03/G.E/2016

جامعة محمد خيضر بسكرة
كلية العلوم و التكنولوجيا
قسم: الهندسة الكهربائية
المرجع:Ref/03/G.E/2016.

Thèse présentée en vue de l'obtention du diplôme de

Doctorat en Sciences

Spécialité: Génie électrique

Option: Electronique

Reconnaissance Automatique du Locuteur à Travers les Canaux Digitaux

Présentée par :

AJGOU Riadh

Soutenue publiquement le 14/02/2016

Devant le jury composé de :

Dr. Okba KAZAR	**Professeur**	**Président**	**Université de Biskra**
Dr. Salim SBAA	**Maitre de conférences 'A'**	**Rapporteur**	**Université de Biskra**
Dr. Abdelmalik TALEB AHMED	**Professeur**	**Examinateur**	**Université de Valenciennes- France**
Dr. Zine-Eddine BAARIR	**Professeur**	**Examinateur**	**Université de Biskra**
Dr. Nabil BENOUDJIT	**Professeur**	**Examinateur**	**Université de Batna**
Dr. Moussa BENYOUCEF	**Professeur**	**Examinateur**	**Université de Batna**

Remerciements

Une thèse n'est pas seulement l'aboutissement d'un travail du doctorant, c'est également une charge pour le jury et les proches. Cette courte page de remerciements leurs est dédiée.

C'est avec émotion que je tiens à remercier tous ceux qui m'ont aidé à élaborer ce travail.

Je tiens tout d'abord à remercier tous les membres de mon jury pour leur présence et leur participation à la soutenance de cette thèse. Mes premiers remerciements à Monsieur ***Okba KAZAR****, Professeur à l'Université de Biskra, qui m'a fait l'honneur de présider la commission d'examen.*

Toute ma gratitude à Monsieur ***Nabil BENOUDJIT*** *et Monsieur* ***Moussa BENYOUCEF*** *Professeurs à l'Université de Batna, tous sont venus de loin juste pour le plaisir de participer à cette journée et acceptent la charge d'être examinateurs de ce travail, j'étais très honoré par leurs présence et d'avoir jugé sans hésitation ce travail ainsi pour ses précieux conseils*

Un grand merci à Monsieur ***Zine-Eddine BAARIR****, professeur au sein de département de génie électrique de l'Université de Biskra qui trouve mes profondes gratitudes pour ses conseils et d'avoir accepté de juger ce travail.*

Je voudrais également remercier le Professeur ***Abdelmalik TALEB AHMED*** *le Co-directeur de cette thèse mais aussi et surtout pour sa gentillesse à toute épreuve, et de m'avoir accueilli au sein du Laboratoire LAMIH de l'Université de Valenciennes pendant neuf mois, un plaisir de travailler avec lui.*

Enfin, bien sure, je dois exprimer mes plus grands remerciements à Monsieur ***Salim SBAA*** *le Directeur de thèse de m'avoir supporté et de m'encourager (à tous les sens du terme) à fond durant toutes ces années et m'avoir permis d'améliorer mes capacités scientifiques et m'a introduit au domaine du traitement de la parole ainsi que de m'avoir donné un bon état d'esprit pour la recherche et de dépasser tous les obstacles mais aussi pour n'avoir cessé de m'assurer son assistance tout au long de ma thèse, en me poussant à continuer dans les moments difficiles. Pour cela, il a toute ma gratitude Merci Dr SBAA Salim.*

Dédicace

JE DÉDIE CE TRAVAIL À

MA GRANDE FAMILLE, MES PARENTS, MES FRÈRES, MES SŒURS

LA MÉMOIRE DE MES GRANDS PARENTS

LA MÉMOIRE DE MON BEAU PÈRE ABBES GUETTAL

MA PETITE FAMILLE, À MON ÉPOUSE DE M'AVOIR SUPPORTÉ PENDANTS TOUTE CES ANNÉES D'ÉTUDE ET ENFIN À MON PETIT GARÇON AHMED SARI

Résumé

Notre travail appartient au domaine du traitement de la parole, précisément la Reconnaissance Automatique du Locuteur (RAL) à travers les canaux digitaux motivé par le développement diligent des réseaux dans le sens large. La reconnaissance automatique du locuteur regroupe les problèmes relatifs à l'identification et la vérification du locuteur sur la base de l'information contenu dans le signal acoustique. À travers la recherche s'est avéré être que l'étape d'identification est l'étape essentielle dans la reconnaissance du locuteur, nous sommes devenus donc plus intéressés à identifier le locuteur que la vérification dans ce travail.

L'objectif final d'un système de RAL est la communication homme-machine. Ce moyen naturel d'interaction a trouvé de nombreuses applications en raison du développement rapide des différents matériels et logicielles, les plus importants sont l'accès aux systèmes d'information, d'aide aux handicapés, ou le contrôle de système à distance.

Ce travail s'agit de reconnaître une personne à partir de sa voix à distance *(remote speaker recognition RSR)*. Ce type d'applications (RSR) a été renforcé clairement par le développement rapide des réseaux numériques (cellulaire et internet) et concentré sur une classe large d'applications qui impliquent l'accès à travers la parole aux systèmes de l'information éloignés c'est à dire à travers IP (Internet Protocol) et les réseaux cellulaires (GSM, UMTS, LTE….). Dans ce sens, on a estimé un système de reconnaissance du locuteur à distance suffisamment robuste en développant un algorithme d'extraction des paramètres (basé sur les paramètres autorégressive (AR) et les coefficients cepstraux (MFCC)) et un algorithme de détection Parole/Silence. D'autre part, on tient compte des effets des problèmes que peut subir les canaux de transmission dans un environnement bruité sur les systèmes de reconnaissance à distance.

Mots clés : reconnaissance automatique du locuteur à distance, réseau internet, réseau mobile, VQ, GMM, MFCCAR, SAD.

Abstract

Our work is in the field of speech processing, specifically the automatic speaker recognition (ASR) through digital channels motivated by the rapid development of networks in the the broad sense. Automatic speaker recognition includes issues related to the identification and speaker verification on the basis of information contained in the acoustic signal. Through research turned out to be that the identification step is the essential step in speaker recognition, we have become thus more interested in identifying the speaker in this work. The ultimate goal of a ASR system is the human-machine communication. This natural way of interaction has found many applications due to the rapid development of various hardware and software, the most important are the access to information systems for the disabled, or control of remote system.

This work is about recognizing a person from its remote voice (remote speaker recognition RSR), this type of application was clearly boosted by the rapid development of digital networks (mobile and internet).The present work is focused on a wide class of applications that require access through the voice to the remote information systems, it means through Internet Protocol (IP) and cellular networks (GSM, UMTS, LTE….)……. Where, we have developped a sufficiently robust remote speaker recognition system by developing a feauture extraction algorithm (based on autoregressive model (AR) and mel-fréquency cepstral coefficients (MFCC)) and a speech activity detection algorithm (SAD). Otherwise, taking into account the problems that can occur transmission channels in a noisy environment and its affects on RSR systems.

Mots clés : Remote speaker recognition (RSR), réseau internet, réseau mobile, VQ, GMM, MFCCAR, SAD.

ملخص

عملنا هو في مجال معالجة الكلام، وتحديدا التعرف التلقائي على المتحدث من خلال قنوات رقمية بدافع من التطور السريع للشبكات بالمعنى الواسع. من خلال البحث اتضح أن خطوة تحديد المتكلم هي خطوة أساسية في التعرف على المتكلمين فأصبحنا مهتمين اكثر بتحديد المتكلم في هذا العمل. الهدف النهائي لنظام التعرف التلقائي على المتحدث هو التواصل بين الإنسان والآلة المبرمجة وقد وجدت هذه الطريقة العديد من التطبيقات وذلك بسبب التطور السريع في مختلف الأجهزة والبرمجيات، وأهمها الوصول إلى نظم المعلومات لذوي الاحتياجات الخاصة، أو السيطرة على النظام عن بعد .هذا العمل هو عن التعرف على الأشخاص عن بعد من خلال الصوت, يركز العمل الحالي على فئة واسعة من التطبيقات التي تتطلب الوصول من خلال الكلمة لنظم المعلومات عن بعد من خلال بروتوكول الإنترنت والشبكات الخلوية. هذا النوع من التطبيق تعزز بوضوح من خلال تطور الشبكات الرقمية (الإنترنت و الهاتف النقال) .في هذا العمل،تم عمل نظام التعرف على المتحدث عن بعد قوي بما فيه الكفاية من خلال وضع خوارزمية جديدة لاستخراج خصائص الناطق وخوارزمية كشف الكلام/السكوت مع الأخذ بعين الاعتبار المشاكل التي يمكن أن تحدثها قنوات البث في بيئة ضجيج على نظام التعرف على المتكلم عن بعد .

الكلمات الشائعة : تحديد هوية الناطق عن بعد ، الانترنت ، شبكة للهاتف المحمول ، GMM، VQ ، MFCCAR، SAD.

Table des matières

CHAPITRE III : Codage de la parole et les effets sur le system de RSR.

CHAPITRE IV : Développement et évaluation d'un système de RSR.

Liste des figures

Liste des tableaux

Liste des symboles et abréviations

RAL	Reconnaissance Automatique du Locuteur
IAL	Identification Automatique du Locuteur
RAP	Reconnaissance Automatique de la Parole (RAP)
VAL	Vérification Automatique du Locuteur
UBM	Universal Background Model
FFT	Fast Fourier Transforme
DCT	Discret Cosin Transform
MFCC	Mel Frequency Cepstral Coefficient
PLP	Perceptual Linear Prediction
AR	Autoregressive
SAD	Speech Activity Detection
VAD	Voice Activity Detection
CMVN	Cepstral Mean and Variance Normalization
CMS	Cepstral Mean Subtraction
DTW	Dynamic Time Warping
VQ	Vector Quantisation
HMM	Hidden Markov Models
SVM	Support Vector Machine
GMM	Gaussian Mixture Models
UBM	Universel Background Model
EM	Expectation-Maximisation
MAP	Maximum A Posteriori
RSR	Remote Speaker Recognition
NSR	Network Speaker/speech Recognition
DSR	Distributed Speech Recognition -DSR
VoIP	Voice over Internet protocol
QoS	Quality of Services
CEPT	Conférence européenne des postes et télécommunications

GSM	Global System of Mobile communication
ETSI	European Telecommunications Standards Institute
IMT	International Mobile Telecommunications
CDMA	Code Division Multiple Access
UMTS	Universal Mobile Telephone System
UTRA	UMTS Terrestrial Radio Access
FDD	Frequency Division Duplex
W-CDMA	Wideband CDMA
TDMA	Time Division Multiple Accès
ITU-T	International Telecommunications Union - Telecoms
DSSS	Direct Sequence Spread Spectrum
OFDMA	Orthogonal Frequency Division Multiple Access
AMRF	Accès Multiple par Répartition en Fréquence
AMRT	Accès Multiple à Répartition dans le Temps
SC-FDMA	Single-Carrier Frequency Division Multiple Access
LTE	Long Term Evolution
AWGN	Additive White Gaussian Noise
LOS	Line-Of-Sight
SNR	Signal to Noise Ratio
ML	Maximum Likelihood Decoding
BPSK	Binary phase shift keying
QPSK	Quadrature Phase Shift Keying
GMSK	Gaussian minimum-shift keying
ARPANET	Advanced Research Projects Agency Network
DARPA	Defense Advanced Research Projects Agency
IETF	Internet Engineering Task Force
RTP	Real Time Protocol
PAPS	Premier Arrivé Premier Servi
FCFS	First Come First-Serve
WFQ	Weighted Fair Queuing
DCT	Discret Cosinus Transform
FEC	Forward error correction
CE	Error Concealment
MOS	Mean Opinion Score

PESQ	Perceptual Estimation of Speech Quality
SAD	Speech Activity Detection
MIC	Modulation par Impulsions Codées
PCM	Pulse Code Modulation
MICD	Modulation par Impulsion Codée Différentielle
DPCM	Differential pulse-code modulation
ADPCM	Adaptive Differential Pulse Code Modulation
QMF	Quadrature Mirror Filters
MELP	Mixed Excitation Linear Prediction
AMDF	Average Magnitude Difference Function
ASDF	Average Square Difference Function
SIFT	Simplified Inverse Filtering
HPS	Harmonic Product Spectrum
CELP	Code Excited Linear Prediction
ACELP	Algébrique CELP
VSELP	Vector Sum Excited Linear Prediction
MMSE	Minimum Mean Square Error
MSE	Mean Square Error
RPE	Regular Pulse Excitation
EVRC	Enhanced Variable Rate Coder
GSM-HR	GSM-Half Rate
GSM-EFR	GSM-Enhanced Full Rate
GSM-AMR	Adaptive Multirate
STP	Short Term Predictor
LTP	Long Term Predictor
QCELP	Qualcomm Code Excited Linear Prediction
EZR	Energy and Zero crossing Rate
EER	Taux d'Egal Erreur
DCF	Fonction de coût de décision
DET	Evolution des deux types d'erreur
DS-CDMA	L'étalement de spectre en séquence directe.

Introduction générale

Ce travail de thèse s'intéresse à la reconnaissance automatique du locuteur (RAL) dans les réseaux de communication (internet, mobiles, …). Plus précisément, nous nous intéressons la reconnaissance du locuteur à travers les canaux digitaux, ce sujet est très vaste, et nécessite plusieurs études et conceptions, c'est-à-dire le traitement de la parole (extraction des caractéristiques, prétraitement, codage…) et télécommunication (réseaux mobiles, réseaux de communication…), dont on a essayé de donner le mieux à travers ce travail.

Le développement rapide des réseaux numériques au cours des dernières années a ouvert un nouveau champ d'expansion pour les techniques vocales. L'objectif final d'un système de RAL est la communication homme-machine. Ce moyen naturel d'interaction a trouvé de nombreuses applications en raison du développement rapide des différents matériels et logicielles. Les plus importants sont l'accès aux systèmes d'information, d'aide aux handicapés, ou le contrôle du système à distance, c'est-à-dire la reconnaissance automatique du locuteur à distance (Remote Speaker Recognition - RSR). La principale différence entre les systèmes RAL et RSR est que RSR implique un réseau numérique placé entre l'utilisateur et le moteur de reconnaissance. En ce sens, on peut considérer que le réseau tels que les téléphonies mobiles, devient une étape supplémentaire du système de RAL.

Problématique et motivation

Le système implique, à son tour, que l'utilisateur peut accéder au système de reconnaissance (RSR) dans un environnement défavorable où le bruit acoustique peut dégrader sérieusement les performances du système. En effet, ce travail a été motivé clairement par le développement rapide des réseaux numériques (cellulaire et Internet), il regroupe les problèmes relatifs à l'identification et la vérification du locuteur et les effets des problèmes que peut subir un canal de transmission sur un système de reconnaissance, dont le thème, c'est la reconnaissance du locuteur à distance. Il est intéressant d'éclairer que les systèmes RSR dépendent de certaines taches indispensables : techniques d'extraction des paramètres et un algorithme de détection parole /silence.

Afin de faciliter ces conceptions globales, cette thèse fournit les notions de base nécessaires à la reconnaissance du locuteur, le codage et la transmission à travers les canaux numériques. Notre principale motivation était d'organiser et de présenter au lecteur les concepts essentiels et contributions liées à la reconnaissance automatique du locuteur à distance.

Objectif

Les systèmes de RSR dont nous nous occupons exigeant un réseau numérique pour leur déploiement. D'habitude, cela sera un réseau de téléphonie mobile ou un réseau IP (Internet protocole). Nous introduisons les traits fondamentaux de ces réseaux qui sont essentiels pour le développement de RSR. Les caractéristiques de reconnaissance (extraction des paramètres de signal parole) sont extraites des signaux reconstruits après avoir été transmis à travers le réseau. Notre objectif dans ce travail est de développer un système de reconnaissance automatique du locuteur à distance (RSR). Bien que le défi majeur en RAL réside dans les taches d'extraction des paramètres, et la détection des zones d'activité vocales, dans ce travail et pour augmenter le taux de reconnaissance, une amélioration d'algorithmes de détection d'activité vocale sera notre cible. Il est ainsi nécessaire de développer une technique d'extraction des caractéristiques de signal parole assez efficace et robuste.

Comme mentionné précédemment, un système de RSR diffère d'un système classique de RAL, les systèmes de RSR sont mis en œuvre sur les réseaux numériques. On se concentre alors sur les architectures de RSR à travers les réseaux numériques. Dont, nous nous focalisons sur les dégradations des systèmes de reconnaissance RSR dues au codage de la parole, la transmission (en tenant compte des dégradations que peuvent subir les canaux de transmissions) et la prise de son (bruits ambiants ou réverbérations).

Contribution

Dans ce travail, nous proposons un système de reconnaissance automatique du locuteur à distance (RSR), dont il y a deux schémas principaux, l'un est un «schéma d'un système de reconnaissance du locuteur dans le réseau» et l'autre est «un schéma d'un système de reconnaissance du locuteur distribué». Nous avons adopté le schéma qui correspond au système RSR dans le réseau. Les thèmes majeurs développés à travers cette thèse sont :

une étude comparative des codeurs/décodeurs de la parole suivant leurs effets sur le taux de reconnaissance sur le système RSR. Cette contribution a fait l'objet d'une communication internationale et un article (annexe A).

- un développement d'un algorithme de détection parole/silence et étudiant leur robustesse en présence du bruit. Cette contribution a fait l'objet d'une communication internationale et un article (annexe A).
- développement d'une nouvelle technique d'extraction des paramètres basant sur les paramètres MFCC et autorégressif (AR). Cette contribution a fait l'objet d'une communication internationale et un article (annexe A).
- une étude comparative des approches de rehaussement de signal parole dégradée par les bruits ambiants ou réverbérations et leurs effets sur notre système RSR.

Organisation de la thèse

La reconnaissance automatique du locuteur regroupe les problèmes relatifs à l'identification et la vérification du locuteur sur la base de l'information contenue dans le signal acoustique. À travers la recherche, il s'est avéré que l'étape d'identification est l'étape essentielle dans la reconnaissance du locuteur. Nous sommes devenus donc plus intéressés à identifier le locuteur dans ce travail.

Cette thèse est constituée d'une introduction générale, quatre chapitres et une conclusion générale et perspectives. Les deux premiers chapitres sont des chapitres d'état de l'art alors que les deux derniers chapitres sont des contributions. La thèse est donc, organisée comme suit :

Le premier chapitre est consacré à la présentation des caractéristiques et modélisations du signal parole ainsi que l'exploration des techniques de reconnaissance automatique du locuteur. On explore alors les outils nécessaires à la reconnaissance automatique du locuteur comme le prétraitement, les techniques d'extraction des paramètres et les différents modèles à savoir : QV, GMM, HMM, GMMHMM, SVM.

Bien que, les systèmes de RSR dont nous nous occupons exigent un réseau numérique pour leur déploiement. Dans le deuxième chapitre, on se concentre sur une large étude des réseaux sans-fil, mobile et Internet. On introduit les traits fondamentaux de ces réseaux qui sont essentiels pour le développement de RSR. On donne un aperçu aux canaux de transmission

utilisés par ces réseaux et la dégradation qu'ils habituellement subissent, et cela pour un objectif d'extraction des paramètres de signal parole après avoir été transmis. D'autre part, on explore les différentes architectures de RSR disponibles dans la littérature.

Dans le troisième chapitre, on parcourt les différentes techniques de codage du signal parole et le codage de source. Ainsi, on propose notre architecture de RSR et étudiant les effets des codecs, en tenant compte de trois types de codec de la parole : PCM, DPCM et ADPCM conformément à la norme de ITU-T (International Telecommunications Union - Telecoms) utilisés en téléphonie et VoIP (Voice over Internet Protocol). Afin d'améliorer les performances de la reconnaissance du locuteur dans un environnement bruité, nous proposons un nouvel algorithme de détection d'activité de la parole (Speech Activity Detection-SAD). Les résultats de ce chapitre ont fait l'objet d'une communication internationale et un article (annexe A).

Dans le quatrième chapitre, nous avons développé un système de reconnaissance du locuteur à distance (RSR) à travers le canal AWGN fondé sur une nouvelle technique d'extraction des paramètres. Cette dernière repose sur la combinaison des paramètres d'autorégressive (AR) et les coefficients cepstraux (MFCC) qui s'avère plus robuste en milieu bruité. Pour améliorer le taux de reconnaissance, une amélioration d'algorithmes de détection d'activité vocale (SAD) vue dans le chapitre III est alors faite en tenant en compte d'estimation du bruit avant (Prior SNR estimation) la décision de parole/non-parole. Ces résultats ont fait l'objet de deux communications internationales et un article (annexe A). En d'autres termes, on a étudié l'effet de deux techniques d'accès multiple nécessaire aux réseaux mobiles et IP à savoir OFDM (multiplexage par répartition orthogonale de la fréquence), DS-CDMA (L'étalement de spectre en séquence directe) sur notre système de RAL à distance. Afin d'améliorer au mieux le taux de reconnaissance en présence du bruit, on a fait une étude comparative des techniques de rehaussement de signal parole et de les appliquer sur notre système de reconnaissance.

On termine ce travail par une conclusion générale et perspectives. Une phase de validation en conditions réelles de fonctionnement est encore nécessaire. Le domaine des techniques d'extraction des paramètres du signal parole reste un sujet important pour les chercheurs. Dans le domaine du taux de reconnaissance, des techniques de rehaussement du signal parole sont nécessaires à utiliser donc nous proposons de développer d'autres techniques de rehaussement.

Chapitre 1

Reconnaissance automatique du locuteur

Sommaire

I.1 Introduction :

La reconnaissance automatique du Locuteur – RAL - s'inscrit dans le domaine du traitement de la parole [1] dont, la figure I.1 présente les différentes taches du traitement de la parole. La reconnaissance automatique du locuteur (RAL) consiste à reconnaitre l'identité d'une personne par l'analyse de sa voix [2]. Objet d'un intérêt accru depuis quelque temps au même titre que l'ensemble des méthodes d'authentification dites biométriques, elle ne figure pas parmi les plus fiables de ces techniques, au premier rang desquelles on retrouve l'analyse des empreintes digitales et génétiques. Cependant la RAL présente un certain nombre de qualité qui la distingue de ces dernières notamment en matière de facilité de déploiement. Tout d'abord, le mode opératoire, un simple enregistrement audio, permet une acceptation plus aisée de la part des utilisateurs par rapport à d'autres techniques d'identifications plus intrusives (notamment du fait que la reconnaissance du locuteur ne requiert aucun contact physique). De même le cout du matériel impliqué est plus réduit [3]. Enfin, la RAL offre l'unique avantage d'être utilisable à distance, sans nécessiter d'autre terminal qu'un simple téléphone [3]. Les caractéristiques de la reconnaissance du locuteur lui ouvrent d'autres champs applicatifs que la simple authentification d'utilisateurs, c'est l'accès à certaines applications à distance qui sera le sujet de notre thèse.

Cependant, le principe de la RAL induit un certain nombre de difficultés auxquelles il faut faire face lors de la mise en œuvre d'un système de reconnaissance du locuteur. En effet la capacité à identifier les locuteurs repose sur les différences entre les voix de divers locuteurs. Mais cette variabilité interlocuteurs se retrouve en concurrence avec la variabilité intra-locuteur (changement de la voix d'un même locuteur entre deux enregistrements), la variabilité de l'environnement d'opération (bruit, niveau d'enregistrement) et du canal de transmission du signal de parole (par exemple lors d'une transmission par téléphone) [3].

Dans ce premier chapitre on explore les outils nécessaires à la reconnaissance automatique du locuteur dans le sens large, ses outils comme le prétraitement de signal parole et techniques d'extraction de paramètres et un état de l'art des différents modèles utilisés dans la littérature (QV, GMM, HMM, GMMHMM, SVM).

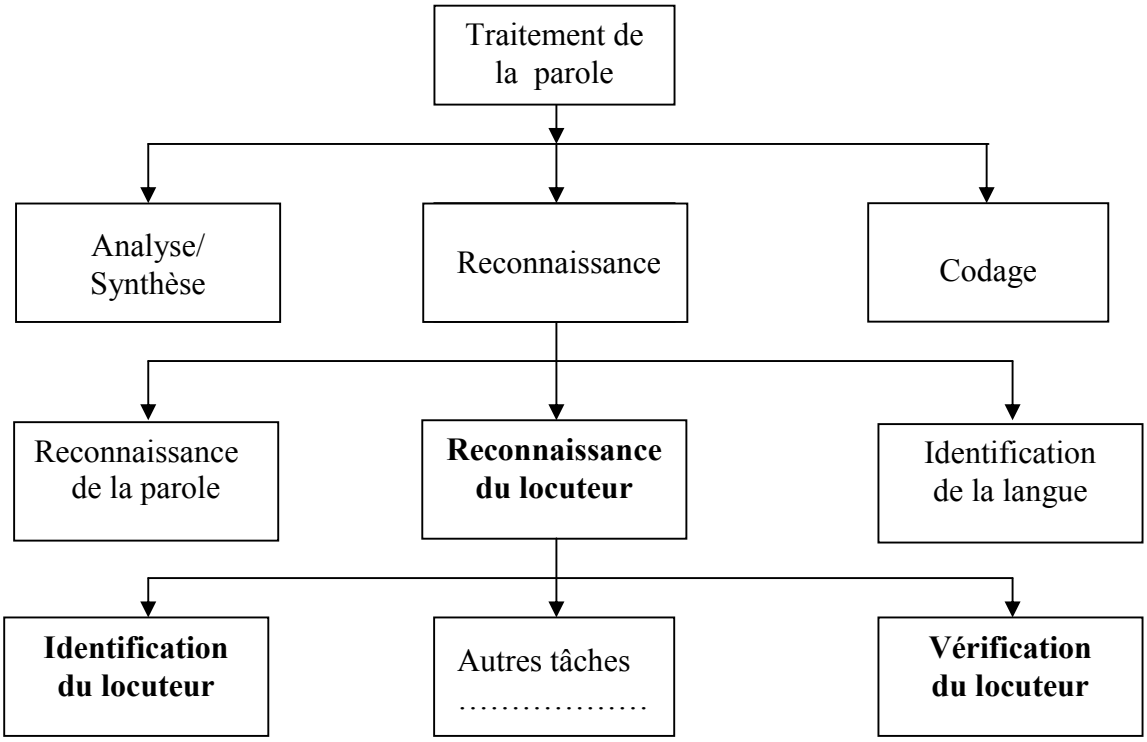


Figure I.1 Différentes taches du traitement de la parole.

I.2 Différentes Tâches en RAL et ses applications

Le plus évident type d'application qui apparaît pour la reconnaissance automatique du locuteur est l'authentification de l'utilisateur au sein d'un système de sécurité comme le cas du contrôle d'accès à un bâtiment, un réseau ou toute autre ressource sensible. Ainsi des applications policières telles que l'automatisation d'écoute téléphonique. L'identification automatique du locuteur [4] et la vérification automatique du locuteur [5] sont les tâches essentielles de la RAL. Plus récemment, les besoins applicatifs ont fait naître de nouvelles tâches comme l'indexation par locuteur [6] de flux audio ou le suivi du locuteur (Speaker Tracking) [7] ou de nouvelles variantes telles que la détection de l'interaction d'un locuteur dans une conversation

I.2.1 Identification automatique du locuteur

Le principe de l'identification automatique du locuteur l'IAL, illustré par la figure I.2, consiste à retrouver l'identité du locuteur associé parmi une population de locuteurs connus. D'un point de vue schématique, une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est " comparée " à une référence caractéristique du locuteur. L'identité du locuteur dont la référence est la plus "proche" de la

séquence de parole est donnée en sortie du système d'IAL [8]. Deux modes sont proposés en IAL : l'identification en ensemble fermé, dont on suppose que la séquence de parole est effectivement prononcée par un locuteur connu et l'identification en ensemble ouvert pour lequel le locuteur peut ne pas être connu. En mode "ensemble ouvert", le système doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée [8]. De par son principe - déterminer une identité parmi des identités potentielles – les performances des systèmes d'IAL se dégradent à mesure que la population de locuteurs augmente [9,10].

I.2.1.1 Applications

En IAL, les applications sont peu nombreuses. On peut retenir, par exemple, l'utilisation d'un système d'IAL en vue de faciliter l'adaptation au locuteur des systèmes de reconnaissance automatique de la parole (RAP). Par ailleurs, il peut être intéressant pour des applications commerciales d'associer un même mot de passe pour une petite population de locuteurs (membres d'une famille, d'une société). Dans une telle situation, un système d'IAL en ensemble ouvert et dépendant du texte peut être utilisé pour contrôler l'accès à des données sensibles [8].

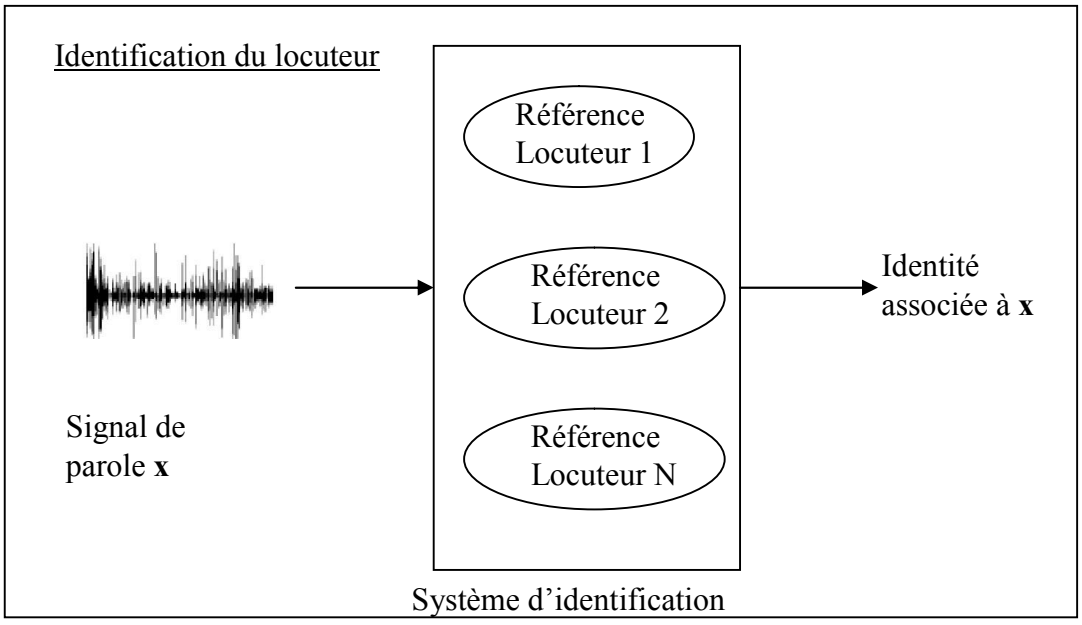


Figure I.2 Principe de base de l'identification du locuteur [8]

I.2.2 Vérification Automatique du Locuteur

La Vérification Automatique du Locuteur (VAL) est le processus décisionnel permettant de déterminer, au moyen d'un message vocal, la véracité de l'identité revendiquée par un individu dont la figure I.3 représente le principe de VAL [11, 12]. L'identité ainsi que le message vocal

constituent les deux entrées du système de VAL. L'identité, nécessairement connue du système, désigne automatiquement la référence caractéristique d'un locuteur. Une mesure de similarité basée sur le rapport de vraisemblance est calculée entre cette référence et le message vocal puis comparée à un seuil de décision. Dans le cas où la mesure de similarité est supérieure au seuil, l'individu est accepté. Dans le cas contraire, l'individu est considéré comme un imposteur et sera rejeté [11, 12].

I.2.2.1 Applications

Les applications de VAL sont multiples et principalement commerciales :

- Serrures vocales pour le contrôle d'accès à des locaux.
- Authentification pour l'accès à distance.
- Protection de matériel contre le vol (téléphones portables, voitures, etc.) ;

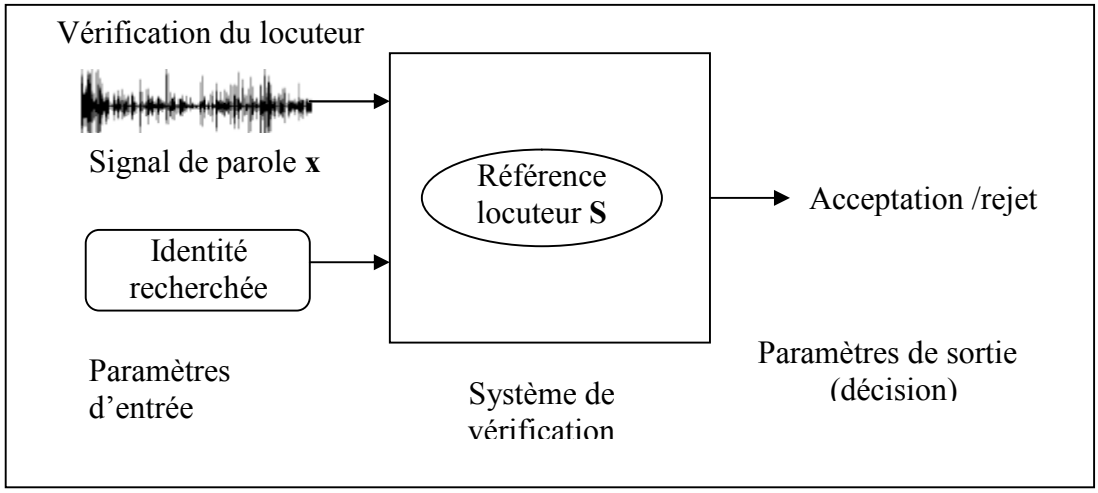


Figure I.3 Vérification Automatique du locuteur [12].

I.2.3 Détection de locuteur dans un flux multi-locuteurs.

Il s'agit d'une extension de la VAL à un test en environnement multi-locuteurs. Le principe est, à partir de l'enregistrement de référence d'un locuteur, de déterminer si ce locuteur est présent au sein d'un enregistrement multi-locuteurs, par exemple une conversation [12].

I.2.4 Suivi de locuteur

Le suivi de locuteur consiste à trouver les limites des interventions du locuteur qu'on a recherché au sein du document multilocuteurs. Il s'agit donc de déterminer si ce locuteur intervient et si oui, quand. La figure I.4 donne une illustration de ce principe.

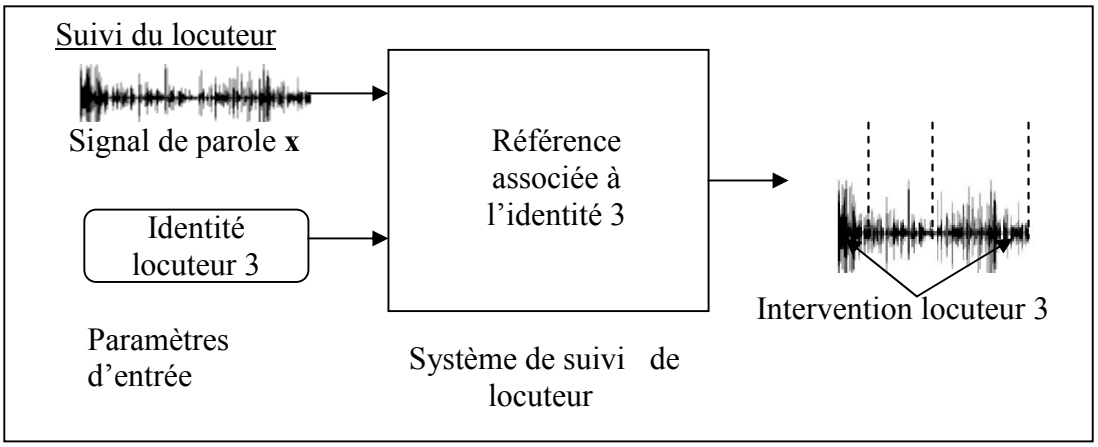


Figure I.4 Tâche de suivi de locuteurs.

I.2.5 Segmentation en locuteurs

C'est la détermination du nombre de locuteurs présents dans un document audio tout en délimitant leurs interventions. La complication de cette tache résulte du traitement de documents pour lesquels peu ou pas d'informations sont connues a priori. Notamment, pas d'information n'est disponible à la primitive concernant les locuteurs participant dans le document : ni leur nombre, ni leur identité, ni aucun échantillon de leur voix permettant d'avoir une référence. Toutes ces informations doivent être extraites du document étudié [3].

I.2.5.1 Applications

Le domaine d'application est la segmentation automatique d'échanges radio entre pilotes et contrôleurs aériens. Depuis, le champ d'application de la segmentation en locuteurs s'est étendu et cette tache se retrouve intégrée dans le cadre plus vaste de l'indexation en locuteurs de bases de données de documents multimédia. Le spectre des types de documents traités s'en trouve élargi : conversations téléphoniques, enregistrement de journaux télévisés ou radiophoniques, films, enregistrements de réunions [3]. Le type de conditions rejointes (parole plus ou moins spontanée, conditions d'enregistrement variables, nombre d'intervenants....) contribue à faire de la segmentation en locuteurs une tache très complexe.

I.3 Mise en place d'un système de RAL

Un système de RAL pour une application donnée se décompose en deux phases distinctes. La première phase est nécessaire à la construction des références ou modèles de chaque locuteur connu du système de chaque client de l'application. Elle consiste à collecter, auprès de ses

clients, des signaux de parole dits d'apprentissage, lors de sessions d'enrôlement. La seconde phase est la phase de reconnaissance à proprement parler qui consiste, pour un client, à se présenter devant le système de RAL (phase de test) [13].

I.4 Problèmes rencontrés en RAL

Les systèmes de RAL souffert des difficultés liées au domaine applicatif, comme l'utilisation des systèmes dans des conditions difficiles, les tentatives d'imposture, etc;

I.4.1 Variabilité due au locuteur

Le signal parole varie pour un même individu parce que la voix d'une personne peut évoluer entre le début et la fin de la journée. Cette variabilité intra-locuteur est induite par l'évolution naturelle ou volontaire de la voix d'une personne.

I.4.2 Variabilité due au matériel

Cette variabilité est due aux: microphone, combiné téléphonique, ligne de transmission (ex : lignes téléphoniques), convertisseurs. Ces informations apparaissent le plus souvent sous la forme de déformations/dégradations du signal de parole.

I.4.3 Robustesse en environnements et tentatives d'imposture

Les systèmes de RAL doivent être robuste face au bruit ambiant et les environnements des canaux digitaux (téléphone, réseaux mobile, internet...). Dans le chapitre suivant on évoquera les réseaux et ses dégradations pour un objectif de développer un system de RAL à travers les canaux de transmission. Un système de RAL peut faire l'objet d'attaques d'individus envahissant l'identité de quelqu'un d'autre. Ces attaques peuvent, par exemple, avoir pour dessein des transactions frauduleuses sur le compte bancaire d'un client ou l'accès à des données confidentielles. Un système de RAL doit par conséquent être robuste.

I.5 System dépendance et indépendant du texte

Diverses applications reposant sur une même tache peuvent se différencier entre autres par leur degré de dépendance au texte. Les systèmes de RAL dits indépendants du texte si ne tiennent aucun compte du contenu linguistique du signal de parole. À l'opposé, les systèmes

dits dépendants du texte s'ils utilisent la connaissance de tout ou partie de ce contenu linguistique pour affiner la reconnaissance du locuteur.

I.6 Les outils de la reconnaissance automatique du locuteur

Un système d'IAL est basé sur la connaissance de "N" clients d'un système, montrés chacun par un modèle. À l'arrivée d'un signal de parole, le système doit déterminer l'identité de la personne qui parle, parmi les N connut. Un système de vérification répond à une autre question, en se basant sur la connaissance du modèle d'une identité clamée "l'et d'un modèle du monde (UBM- Universal background Model), qui représente en réalité l'hypothèse opposée de production. Le système amène si le locuteur "I" parle ou non dans l'enregistrement actuel.

La majorité des systèmes de reconnaissance du locuteur que ce soit dans les taches d'identification ou de vérification, utilisent les modèles de mélange de lois Gaussiennes (Gaussian Mixture Models - GMM) dans la modélisation des locuteurs, que ce soit exclusivement ou en combinaison avec d'autres techniques comme HMM (Hidden Markov Model) ou SVM (Support Vecteur Machine).

Un système de reconnaissance (identification ou vérification) comporte plusieurs composantes : un module d'extraction de paramètres, un bloc d'appariement, un module de normalisation des scores d'appariement et un module de décision. La figure I.5 donne l'architecture d'un système de RAL comprenant l'identification et la vérification.

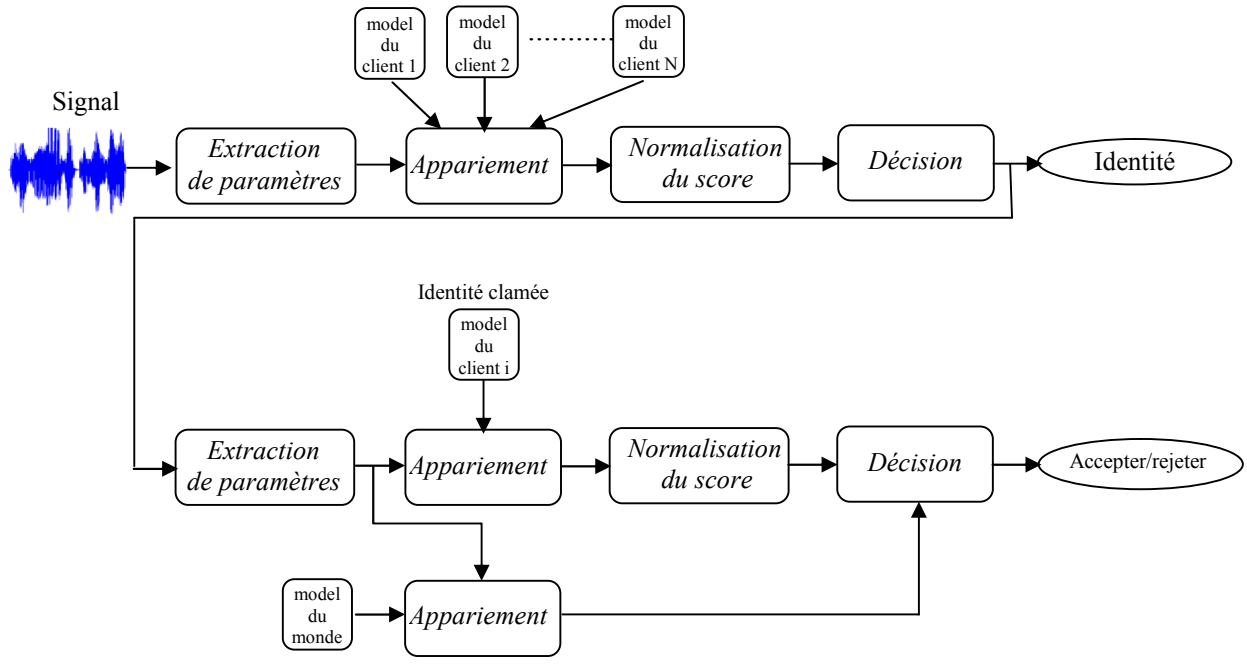


Figure I.5 Schéma typique d'un système de RAL (Identification et Vérification).

Le module d'extraction de paramètres transforme un signal de parole en une séquence de vecteurs acoustiques utiles à la reconnaissance. Ce module comporte différents sous-modules à savoir, la paramétrisation, la segmentation parole / non parole (Speech/non-speech) et des prétraitements. On calcule les log-vraisemblances des vecteurs de paramètres par rapport aux N modèles des clients du système dans une application d'identification. Ces scores d'appariement sont ensuite normalisés pour réduire les effets de la variabilité intersessions, et le module de décision sélectionne le modèle (l'identité) le plus vraisemblable a posteriori. Pour l'application de vérification, on calcule à la fois la log-vraisemblance (normalisés) des vecteurs de paramètres par rapport au modèle de l'identité clamée "I", et par rapport au modèle du monde. Le module de décision compare finalement le rapport des deux scores d'appariement par rapport à un seuil de décision pour déterminer si le locuteur cible "I" parle ou non dans le signal de parole.

Nous présentons par la suite les différentes approches et techniques utilisées en extraction de paramètres ainsi modélisation et normalisation des scores.

I.6.1 Extraction de paramètres

Dans cette sous-section, on explore les différents paramètres du signal parole qui sont utilisés en reconnaissance automatique du locuteur, à savoir des paramètres spectraux, des paramètres liés à la source vocaux, des paramètres prosodiques et des paramètres de haut niveau. Nous parlerons ensuite de la segmentation parole / non parole qui tend à ne garder que les trames utiles au processus de reconnaissance. Bien qu'elle soit faite généralement après la phase de paramétrisation, la segmentation peut aussi intervenir avant celle-ci. Nous exposerons finalement les principales techniques de normalisation des paramètres acoustiques qui ont été proposées en reconnaissance automatique du locuteur [13].

I.6.1.1 Paramètres prosodiques

Les traits ou indices prosodiques d'un signal de parole sont sa fréquence fondamentale (ou pitch), son énergie et son timbre. Le pitch représente la fréquence de vibration des cordes vocales. Cet élément est différent pour la voix d'un homme (entre 60 Hz et 150Hz), la voix d'une femme (aux alentours de 250Hz) ou celle d'un enfant (entre 300Hz et 400Hz). L'énergie d'un son est liée à la pression de l'air en amont du larynx et caractérise son intensité. Le timbre est la caractéristique d'un son permettant de le différencier d'un autre son (c'est la mélodie d'un mot ou d'une phrase).

a) Energie totale :

L'énergie à court terme du signal de la parole fournit une représentation convenable qui reflète ces variations d'amplitude. Elle est calculée à partir de la relation suivante [14]:

$$\overline{E}(m) = \sum_{n=0}^{N-1} x^2(n).w(m-n) \qquad \text{(I.1)}$$

Avec $\overline{E}(m)$ **:** L'énergie moyennes d'une trame "m".

N **:** la largeur de la fenêtre d'analyse.

$x(n)$ **:** C'est le signal de la trame avec une longueur "N".

b) Fréquence fondamental

La période du fondamental est par définition la fréquence de vibration des cordes vocales, elle est appelée aussi le pitch. La figure I.6 représente la fréquence fondamentale.

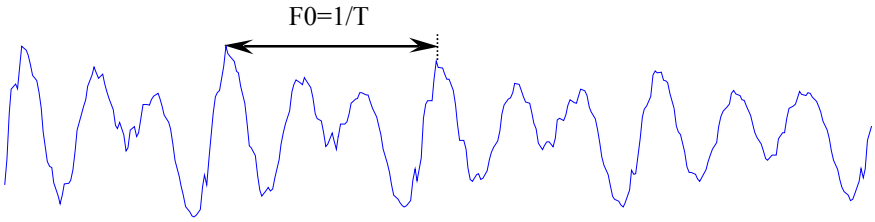


Figure I.6 Fréquence fondamentale**.**

L'extraction du pitch est une tâche particulièrement difficile pour trois raisons :

- La vibration des cordes vocales n'a pas nécessairement une périodicité complète.
- Il est difficile de séparer le pitch des effets du trait vocal.
- La plage de dynamique de la fréquence fondamentale est très grande.

I.6.1.2 Analyse spectrale du signal de parole

L'analyse spectrale de la parole présente des avantages au niveau de la perception, car l'oreille humaine effectue ce genre d'analyse. On introduit en traitement du signal parole les outils suivants :

a) La transformée de Fourier discrète :

L'analyse spectrale se fait à l'aide de la transformée de Fourier. Le signal parole est un signal non stationnaire à long terme mes présumé stationnaire pour une durée allant de 10 à 30

ms. Pour cela on applique le Transformé de Fourier à cours terme (FFT- Fast Fourier Transforme) où il est nécessaire d'effectuer préalablement un fenêtrage s'appel des trames allant de 20 à 30 ms avec un recouvrement ("over lapping" généralement de 30 à 50%). Avec $X(n)$ le spectre du signal numérique x *(k)* [15] :

$$X(n)=\sum_{k=0}^{N-1}x(k)\times e^{-j\pi\frac{nk}{N}} \qquad \text{(I.2)}$$

b) Les paramètres AR

La modélisation d'un signal *x (k)* consiste à lui associer un filtre linéaire qui soumit à une excitation particulière reproduit ce signal le plus fidèlement possible [16]. L'objectif essentiel de la modélisation AR d'un signal est de permettre sa description par un ensemble très limité de paramètres [16] :

$$x(k)+a_1x(k-1)+....+a_px(k-p)=e(k) \qquad \text{(I.3)}$$

Les coefficients de la combinaison linéaire sont trouvés de façon à minimiser l'erreur, Cette modélisation correspond à un modèle tout pôle ; Soit *H (z)* sa fonction de transfert [16] :

$$H(z)=\frac{1}{A(z)} \quad \text{Où} \quad A(z)=1-\sum_{i=1}^{p}a_iz^{-1} \qquad \text{(I.4)}$$

c) Les coefficients de réflexion (ou PARCOR K_i)

Les coefficients K_i calculés par l'algorithme de Levinson [16] caractérisent complètement les coefficients de prédiction a_i et exprime très simplement la stabilité du modèle AR ($|K_i|<1$). Ils sont appelés coefficient de corrélation partielle. Les paramètres a_i et K_i sont appelés couramment paramètres LPC [16]. Le codage LPC traite le signal de parole en faisant une distinction entre les parties voisées et non voisées du signal parole. Les parties voisées présentent une certaine périodicité, ce qui permet de trouver une fréquence fondamentale. Les paramètres renvoyés par cette analyse sont la *fréquence fondamentale* (uniquement pour les parties voisées), *un gain* et les coefficients d'un *filtre tout pôle* (filtre autorégressif AR). En effet, l'analyse LPC est une analyse prédictive qui utilise donc les valeurs entourant un échantillon pour le prédire. Chaque tranche va être passée dans un algorithme afin de calculer les coefficients a_i. Ces coefficients sont ceux du dénominateur du filtre tous pôles dont la transmittance est l'enveloppe spectrale du signal. Le calcul des coefficients se base sur la résolution des équations de **Yule-Walker** qui peuvent notamment être résolues par l'algorithme

de ***Le Vinson*** et l'algorithme de ***Schur*** [17]. Le premier coefficient est toujours égal à un, les suivants sont les a_i qui représentent le dénominateur du filtre (H(z)) autorégressif. (p : l'ordre de paramètre) [17] :

$$H(z) = \frac{\text{gain}}{1 + a_1 z^{-1} + a_2 z^{-2} + ... + a_p z^{-p}} \quad \text{(I.5)}$$

d) Coefficient MFCC

Le signal acoustique contient de différentes sortes de renseignements sur le locuteur. La paramétrisation MFCC (Mel Frequency Cepstral Coefficients) est basé sur la perception humaine de son, sur l'évidence connue que les renseignements portés par les composantes de la fréquence basse du signal de parole sont plus importants phonétiquement pour les humains que les composantes à haute fréquence [17]. La fréquence perceptive humaine est représentée dans l'échelle de Mel qui est l'espacement de fréquence linéaire au-dessous de 1000 Hz et un espacement logarithmique au-dessus de 1000 Hz. On suppose que la perception humaine de son se compose du banc de filtres. Chaque filtre a une forme triangulaire. Les bancs de filtre triangulaires dans l'échelle de Mel sont espacés uniformément.

Après études sur l'oreille humaine, il a été montré que l'homme se base sur une échelle fréquentielle spécifique. La formule de transfert est [4] :

$$mel(f) = 2595 * \log 10\left(1 + \frac{f}{700}\right) \quad \text{(1.6)}$$

f_{low} et f_{high} sont les limites basses et à haute fréquence de banc de filtre, donnés par [13]:

$$f_{low} = \frac{fs}{N} \quad \text{(I.7)}$$

N: est la longueur de trame de 160 échantillons (corresponds à 20ms).

$$f_{high} = \frac{fs}{2} (f_{low} = 100 \text{ Hz}, f_{high} = 8000\text{Hz}) \quad \text{(I.8)}$$

En conséquence à l'équation (I.6) nous avons la figue I.7 qui représente le rapport entre l'échelle de fréquence réelle et son échelle de Mel-fréquence, quand la fréquence réelle *f* est ci-dessous 1000Hz, le rapport est linéaire; cependant, le rapport des traits devient logarithmique quand *f* est au-dessus 1000Hz, la bande passante du filtre individuel (le banc de filtre)

augmente logarithmiquement dans l'échelle normale. Chaque filtre triangulaire une longueur 1000 (Arbitrairement choisit) dans le domaine de fréquence. Notez aussi que le 1000 ème échantillon correspond à : *fs/2* [14]. La Figue I.8 illustre le banc de filtres dans l'échelle de mel-fréquence. (1 dans l'axe X correspond *à fs/2*).

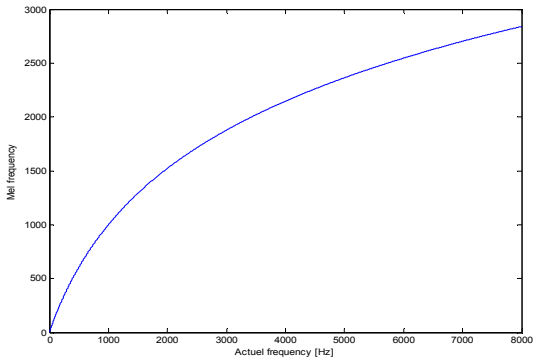


Figure I.7 Rapport entre l'échelle de fréquence réelle et son échelle de Mel-fréquence.

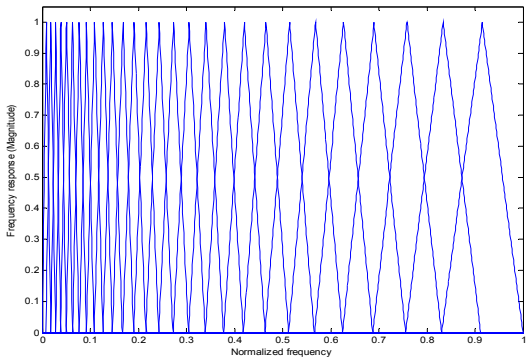


Figure I.8 Banc de filtres dans l'échelle de Mel-fréquence. 1 dans l'axe X correspond à fs/2 : (8000 Hz).

Une fois le spectre mel a été calculé, il doit être converti en arrière à l'intervalle de temps en utilisant le DCT (Discret Cosine Transform). On appelle le résultat des fréquences de mel cepstreaux par les coefficients (MFCCs). En utilisant la même procédure, un ensemble des fréquences de mel sont calculés pour chaque trame de signal parole d'environ 20 millisecondes avec un chevauchement. Le calcul de MFCCs est montré dans la Figue I.9. Le calcul des MFCC se décompose en cinque phases (voir figure I.9) :

- Découper le signal en plusieurs fenêtres avec chevauchement.
- Afin de diminuer la distorsion spectrale souvent on applique une fenêtre de Hamming [17] :

$$W(n) = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N-1}\right) \qquad \text{(I.9)}$$

Cette fonction est multipliée par le signal à transformer, nous minimisons ainsi la distorsion spectrale créée par le chevauchement (le recouvrement).

- Application de la FFT aux fenêtres.
- On passe à l'échelle de Mel (équation I.6).

Pour simuler l'oreille humaine, il faut passer par un banc de Filtres, un filtre pour chaque fréquence que l'on cherche. Ces filtres ont une réponse de bande passante triangulaire. Pour

connaître l'intervalle entre chaque filtre, on utilise une constante : l'intervalle de fréquence de Mel (Mel-Frequency interval).

- Convertissons de spectre logarithmique de Mel en temps comme suist [17] :

$$y(k) = \sum_{n=1}^{N} w(n)\, x(n) \cos\left(\frac{\pi\,(2n-1)(k-1)}{2N}\right) \quad \text{Avec k=1,…., N} \qquad \text{(I.10)}$$

Où : $w(n) = \sqrt{\frac{1}{N}}$ $si\ n = 1\ et\ w(n) = \sqrt{\frac{2}{N}}$ $si\ non$.

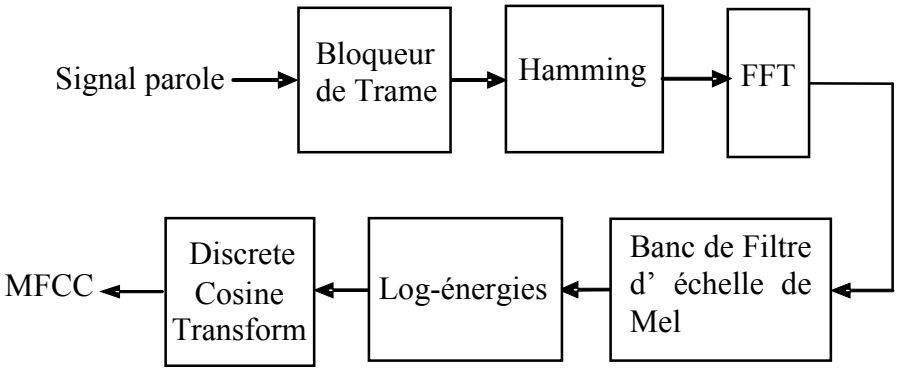


Figure I.9 Calcul des coefficients MFCC avec une échelle Mel.

E) Les coefficients PLP (Perceptuel Linear Predictive).

La méthode de coefficients de prédiction à base de notions psycho acoustiques connue sous le nom PLP (*Perceptual Linear Prediction* ou *Perceptually based Linear Prediction*), est une méthode inspirée du principe de prédiction linéaire (LPC). Elle combine ce principe à une représentation du signal qui suit l'échelle humaine de l'audition. La figure I.10 résume le principe de la méthode dont une analyse spectrale est effectuée au signal parole afin d'obtenir un spectre suivant une échelle d'audition. Ce spectre est ensuite modifié par une interpolation et une transformée de Fourier inverse, le signal obtenu étant passé dans un filtre pour réduire la dimension du spectre et augmenter la résolution fréquentielle [18, 19].

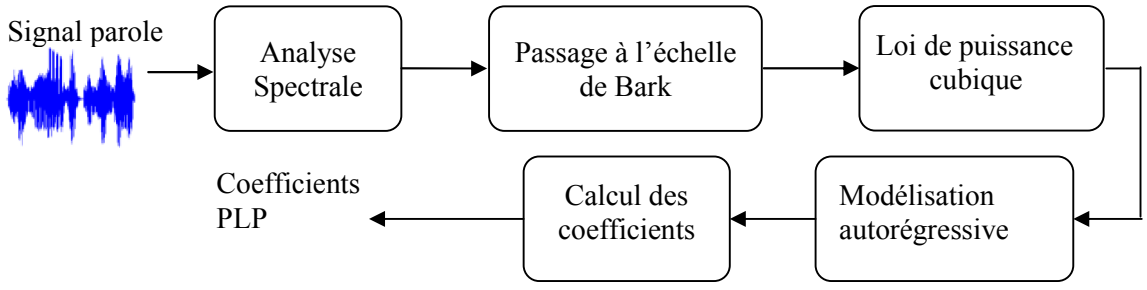


Figure I.10 Méthode de calcul des coefficients PLP.

Dans la première étape on a :

- une analyse en bandes critiques selon une échelle de Bark par un banc de filtres ;
- une préaccentuation des valeurs obtenues selon une courbe suivant approximativement les mêmes principes que les traitements effectués par l'oreille, avec accentuation des basses fréquences et atténuation des hautes fréquences [19] ;

Dans la deuxième étape on a :

- une interpolation des sorties de banc de filtre pour obtenir un spectre sur une échelle fréquentielle auditive,
- une transformée de Fourier inverse.
- une résolution d'un ensemble d'équations linéaires pour obtenir les coefficients issus d'un filtre tout pôle d'ordre 5 (ce qui permet d'obtenir au moins deux sommets caractéristiques selon les auteurs dans [19]).

Cette méthode a pour avantage de permettre une analyse ou un codage de la parole qui respecte le principe de la prédiction linéaire, qui suit l'échelle fréquentielle observable dans l'oreille et enfin, qui réduit l'espace de représentation.

F) Rasta PLP

La méthode PLP [19], dont l'algorithme repose sur des spectres à court terme de la parole, résiste difficilement aux contraintes qui peuvent lui être imposées par la réponse fréquentielle d'un canal de communication. Pour atténuer les effets de distorsions spectrales linéaires, H. Hermansky [20] a proposé de modifier l'algorithme PLP en remplaçant le spectre à court terme par un spectre estimé dont chaque canal fréquentiel est modifié par passage à travers un filtre. Cette modification est à la base de la méthode RASTA PLP. La mise en place de ce filtrage permet, lorsqu'il est effectué dans le domaine spectral logarithmique, de supprimer les composantes spectrales constantes, supprimant ainsi les effets de convolution du canal de communication. Différentes études réalisées avec cette méthode [21,22] ont permis de confirmer les bonnes qualités de cette méthode relativement aux distorsions et ses moindres qualités face aux bruits qualifiés d'additifs, signe de la présence de plusieurs sources sonores dans un même environnement.

Pour améliorer encore la méthode PLP, H. Hermansky définit la méthode J-RASTA, plus résistante aux bruits additifs par adjonction d'un filtrage passe-bas dans le domaine spectral [23].

I.6.1.3 Spectrogrammes

Le spectrogramme permet de mettre en évidence les différentes composantes fréquentielles du signal à un instant donné. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux axes : temps et fréquence. Ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants. La figure I.11 Spectrogramme de mot « Bonjour ».

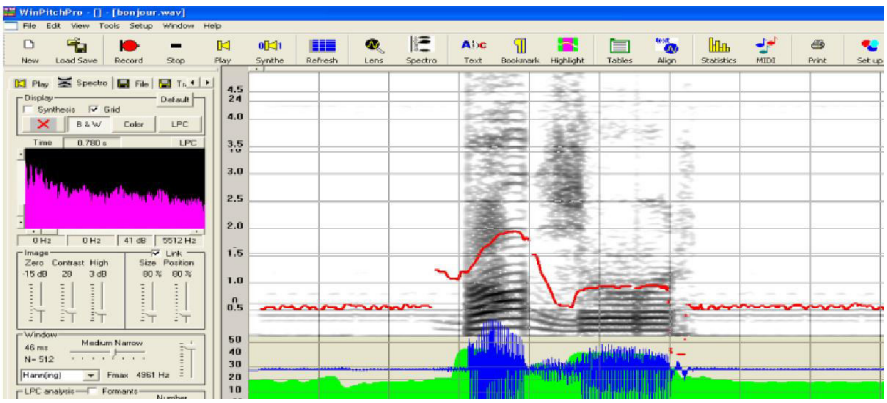


Figure I.11 Spectrogramme de mot « Bonjour », réalisé avec le logiciel WinPitchPro [24].

I.6.1.4 Prétraitement du signal parole

Les zones de silence doit être éliminé. Dans la phase de pré-traitement, les portions de silence sont éliminées des signaux de parole en utilisant un algorithme de détection parole/non-parole (Speech Activity detection- SAD ou Voice Activity Détection-VAD). Il y a plusieurs techniques de détection parole/non-parole dans la littérature [25, 26, 27, 28,29]. Dans notre thèse, on a développé un nouvel algorithme de détection parole/non-parole (voir chapitre IV). L'algorithme SAD s'applique avant l'extraction des paramètres (MFCC, AR, PLP,....) pour que les paramètres soient extraits dans les zones qui expriment réellement la parole. D'autre part, le signal parole est pré-accentué avec un facteur de pré-accentuation "μ"[10]. La préaccentuation consiste à faire passer les tranches de signal dans un filtre passe-haut du premier ordre, de transmittance [23]:

$$T(z) = 1 - \mu \times z^{-1} \qquad \text{(I.11)}$$

Dont, La constante "μ" est le facteur de pré-accentuation souvent égale à 0.97 [23].

Après la phase de fenêtrage, pré-accentuation et détection parole/ non-parole, l'étape d'extraction de paramètres se commence.

I.6.1.5 Normalisation des paramètres acoustiques

Pour augmenter la robustesse des systèmes face aux variations des conditions d'enregistrement et de transmission, différentes techniques de normalisation des paramètres acoustiques ont été proposées en reconnaissance automatique du locuteur. En pratique, différentes techniques de normalisation sont utilisées ensemble dans les systèmes de reconnaissance du locuteur. Parmi ces techniques on a moyenne cepstrale et la normalisation variance (Cepstral Mean and Variance Normalization) CMVN [30], est une technique très simple et très utilisée. Elle consiste à retirer la moyenne de la distribution de chacun des paramètres cepstraux (la composante continue), et à ramener la variance à une variance unitaire en les divisant par l'´ecart type global des paramètres acoustiques [1].

Ainsi, quand seule la moyenne est normalisée, on parle alors de la Soustraction Moyenne Cepstrale (Cepstral Mean Subtraction (CMS) [31]. Parmi les techniques récentes on a le filtrage RASTA [20] comme alternative à la CMS, la "Gaussianization" (caractéristique de déformation ou distorsion) [32] et court délai gaussianization (Short-time Gaussianization) [33].

I.6.2 Modèles de reconnaissance

Les techniques de modélisation et extraction de paramètres sont les parties principales d'un système de reconnaissance du locuteur. Dans cette partie on explore quelque techniques de modélisations utilisés dans la RAL. On distingue l'approche vectorielle, connexionniste, prédictive et statistique.

I.6.2.1 Approche vectorielle

Dans l'approche vectorielle, les vecteurs de paramétriques d'apprentissage et de test sont comparés, sous l'hypothèse que les vecteurs d'une des séquences sont une réalisation imparfaite des vecteurs de l'autre séquence. La distorsion entre les deux séquences représente leur degré de similarité. Cette approche compte deux grandes techniques, e.g., l'alignement temporelle (Dynamic Time Warping) DTW [34] et la quantification vectorielle (Vector Quantisation-VQ) [15, 35, 36], qui ont été respectivement proposés pour les applications dépendantes et indépendantes du texte.

La DTW aligne temporellement les suites d'observations, tandis que la VQ représente le locuteur par un dictionnaire de codes (Codebook) [1].

A) L'alignement temporelle

D'une façon générale, la DTW est une méthode qui recherche un appariement optimal entre deux séries temporelles. L'alignement temporel, plus connu sous l'acronyme de DTW, est une méthode fondée sur un principe de comparaison d'un signal à analyser avec un ensemble de signaux stockés dans une base de référence. Le signal à analyser est comparé avec chacune des références et est classé en fonction de sa proximité avec une des références stockées. Le DTW est en fait une application au domaine de la reconnaissance de la parole. Cette méthode d'alignement de séries temporelles est souvent utilisée dans le contexte de modèles de Markov cachés (HMM) [37].

B) Quantification Vectorielle (VQ)

La quantification vectorielle décompose l'espace acoustique d'un locuteur donné X, en un ensemble de M sous-espaces représentés par leur vecteurs centroides $C = \{c_1, c_2, ...c_M\}$. Ces vecteurs centroides forment un dictionnaire (de taille M) qui modélise ce locuteur, et sont calculés en minimisant l'erreur de quantification moyenne (distorsion) induite par le dictionnaire sur les données d'apprentissage du locuteur $\{x_1, x_2, ...x_M\}$ [1]:

$$D(X,C) = \frac{1}{T}\sum_{t=1}^{T} \min d\left(x_t, c_m\right) \quad , \qquad 1 \le m \le M \tag{I.12}$$

Où $d\left(x_t, c_m\right)$ est une mesure de distance au sens d'une certaine métrique liée à la paramétrisation. L'apprentissage vise à réduire l'erreur de quantification. On peut mieux représenter le locuteur en augmentant la taille du dictionnaire, mais le système sera moins rapide et plus demandeur de mémoire. Il faut trouver donc un bon compromis. La construction du dictionnaire peut être faite en utilisant par exemple l'algorithme LBG [38].

La figure I.10 illustre un schéma conceptuel pour illustrer ce processus de reconnaissance. Dans l'illustration, apparaissent seulement deux locuteurs et deux dimensions de l'espace acoustique. Les cercles désignent les vecteurs acoustiques du locuteur 1 alors que les triangles sont du locuteur 2. Dans la phase de formation (apprentissage), à l'aide de l'algorithme de clustering décrite par la suite, un livre de codes (codebook) VQ spécifique est généré pour chaque locuteur connu par ses vecteurs acoustiques de la formation de cluster. Les mots de code de résultat (le centre de gravité (centroid)) sont indiqués dans la Figure I.12 de cercles

noirs et des triangles noirs pour locuteur 1 et 2, respectivement. La distance entre un vecteur et le mot de code le plus proche d'un livre de codes est appelée une VQ-distorsion. Dans la phase de reconnaissance, on forme le VQ-distortion» à l'aide de chaque codebook d'énoncé d'entrée d'une voix inconnue. Le locuteur correspondant au codebook VQ avec plus petite distorsion totale est identifié comme le locuteur de l'énoncé d'entrée.

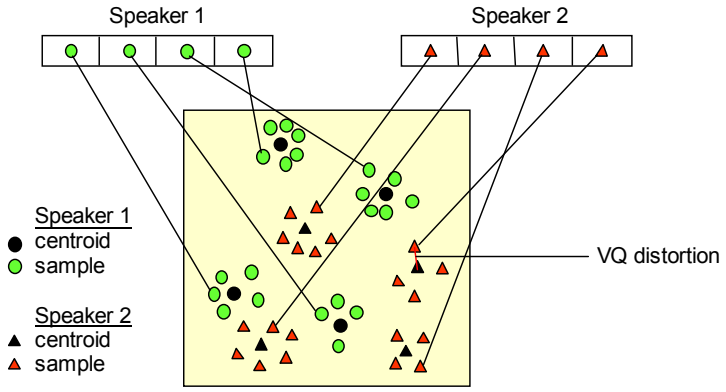


Figure I.12 Diagramme conceptuel illustrant un dictionnaire de codes (codebook) pour le (VQ). Un locuteur peut être distingué sur la base d'une autre de l'emplacement des centroide.

B.1 Cluster les vecteurs d'apprentissage (Clustering the Training Vectors)

Après l'étape d'apprentissage, les vecteurs acoustiques extraits fournissent un ensemble de vecteurs d'apprentissage. La prochaine étape importante est de construire un dictionnaire VQ spécifique de chaque locuteur en utilisant ces vecteurs d'apprentissage par l'algorithme de LBG [37]. Pour cluster un ensemble de vecteurs "L" des vecteur d'apprentissage dans un ensemble de "M" vecteurs de dictionnaire (codebook). L'algorithme est officiellement mis en œuvre par la procédure récursive suivante:

1. Concevoir un livre de codes 1-vecteur; c'est le centre de gravité (centroid)) de l'ensemble de vecteurs d'apprentissage (par conséquent, aucune itération n'est nécessaire ici).

2. Doubler la taille de codebook en divisant chaque codebook en cours y_n selon la règle:

$$y_n^+ = y_n(1+\varepsilon) \tag{I.13}$$

$$y_n^- = y_n(1-\varepsilon) \tag{I.14}$$

Où 'n' varie de 1 à la taille actuelle de la table de codage (codebook), et ε est un paramètre de partage. (Généralement on choisit $\varepsilon = 0.01$).

3. Recherche le proche voisin: pour chaque vecteur d'apprentissage, trouver le mot de code (codeword) dans le livre de code (codebook) actuel qui est la plus proche (en termes de mesure de similarité), et attribuer ce vecteur à la cellule correspondante (associé avec le mot de code le plus proche).

4. Le centroïde à mettre à jour: mettre à jour le mot de code (codeword) dans chaque cellule en utilisant les centroides des vecteurs d'apprentissage affectés à cette cellule.

5. Itération 1: répétez les étapes 3 et 4 jusqu'à ce que la distance moyenne soit inférieure à un seuil prédéterminé

6. Itération 2: répéter les étapes 2, 3 et 4 jusqu'à une taille de bibliothèque de codes de M soit conçu.

Intuitivement, l'algorithme LBG conçoit un codebook de M-vecteurs. On commence d'abord par la conception d'un vecteur de codebook. Puis on utilise une technique de séparation sur les codeword pour initialiser la recherche de 2-vecteurs de codebook, et le processus se continue jusqu'à ce que le fractionnement de la table de codage M-vecteurs désiré soit obtenu. La figure I.13 représente le diagramme avec les étapes détaillées de l'algorithme LBG. "Vecteurs de cluster" est la plus proche voisine procédure de recherche qui attribue chaque vecteur d'apprentissage à un groupe associé au codeword le plus proche [39]. Trouvez les centroïdes est la procédure essentielle. "Calculer D (distorsion)" résume les distances de tous les vecteurs d'apprentissage à la recherche du plus proche voisin afin de déterminer si la procédure a convergé.

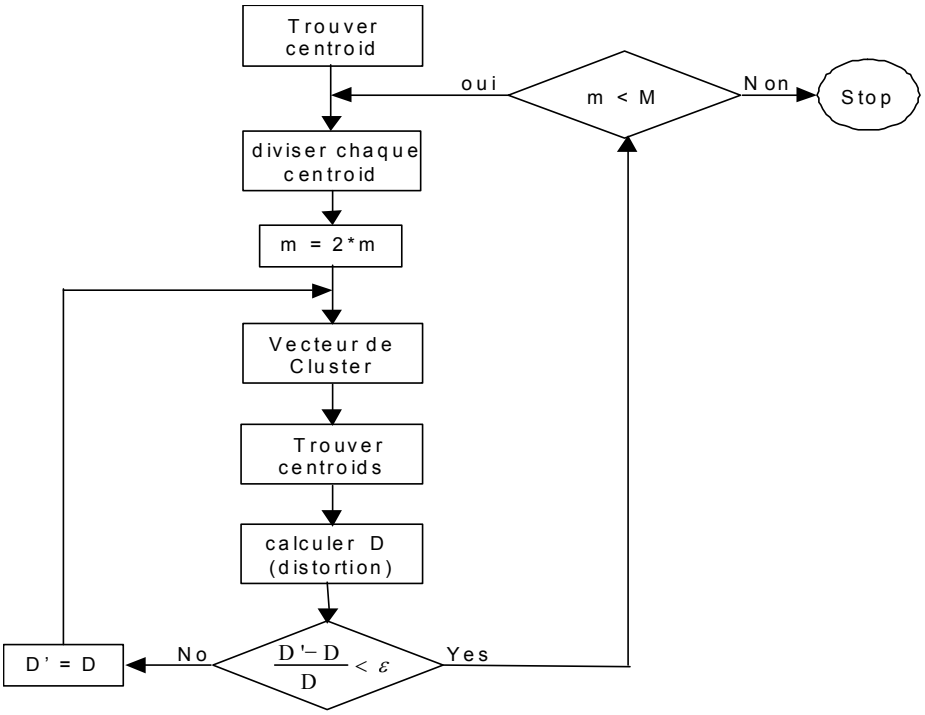


Figure I.13. Diagramme de LBG (Rabiner et Juang, 1993 [39]).

I.6.2.2 Approche prédictive

Les modèles prédictifs reposent sur le principe qu'une trame d'un signal de parole peut être générée à partir des trames précédentes du signal. Un locuteur donné est représenté par une fonction de prédiction estimée sur ses données d'apprentissage. Deux stratégies peuvent être ensuite adoptées pour la reconnaissance : soit calculer une erreur de prédiction comme mesure de similarité, entre les trames et les trames réellement observées ; soit comparer la fonction de prédiction du locuteur concerné avec une nouvelle fonction de prédiction estimée cette fois ci sur les nouvelles données, selon une mesure de distance donnée.

Dans la littérature deux grandes familles de fonctions de prédiction: Les modèles Autorégressifs vectoriels (AR-Vector Models) [15] et Les réseaux de neurones prédictifs [40].

I.6.2.3 Approche connexionniste

Un modèle connexionniste, ou modèle neuromimétique ou réseau de neurones artificiels est un modèle discriminant formé d'un grand nombre de cellules élémentaires (neurones) fortement interconnectées, dont la sortie de chaque neurone est en fonction de ses entrées fortement interconnectées, dont la sortie de chaque neurone est en fonction de ses entrées[41].

I.6.2.4 Approche statistique

Les techniques statistiques considèrent le locuteur comme étant une source probabiliste et le modélisent par une densité de probabilité connue. La phase d'apprentissage consiste à estimer les paramètres de la fonction de densité de probabilité. La décision est prise en calculant la vraisemblance des données par rapport au modèle appris préalablement. Les modèles de Markov cachés HMM [42] ont été utilisés dans des applications dépendantes du texte de reconnaissance automatique du locuteur tandis que les Modèles de Mélange de lois Gaussiennes GMM [43] et les machines à vecteurs de support SVM [44] largement utilisés en indépendant du texte ainsi dans des applications de vérification du locuteur [45,46]. Dans notre thèse on s'intéresse mieux au modèle indépendant du texte (VQ, GMM).

A) Modèle du mélange (GMM)

Un modèle de gaussiennes est une somme pondérée de M densités gaussiennes. La figure I.14 décrire le modèle à mélange gaussiennes. Soit un locuteur s et un vecteur acoustique x de dimension D, le mélange de gaussiennes est défini comme suit [43] :

$$p(x/\lambda_s)=\sum_{m=1}^{M}\pi_m^s\, b_m^s(x) \qquad \text{(I.15)}$$

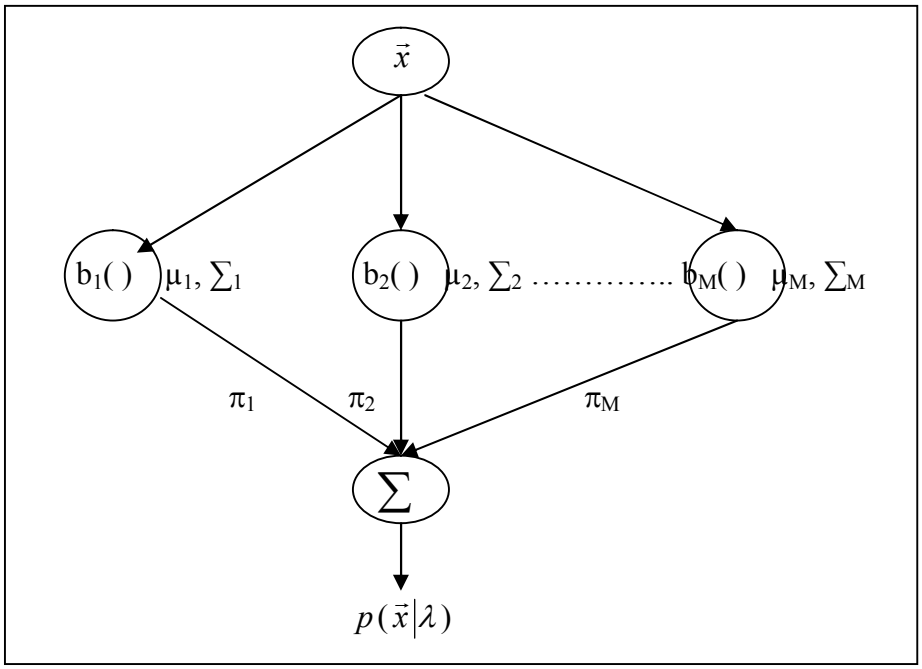


Figure. I.14 Description de modèle à mélange gaussiennes $p(\vec{x}|\lambda)$.

Où les $b_m^s(x)$ représentant des *densités gaussiennes*, paramétrées par un vecteur de moyenne μ_m^s et une matrice de covariance $\sum_m^s$ [43]:

$$b_m^s(x)=\frac{1}{(2\pi)^{D/2}.\left|\sum_m^s\right|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu_m^s)'(\sum_m^s)^{-1}(x-\mu_m^s)\right] \qquad \text{(I.16)}$$

Et les π_m^s représentent les poids du mélange, avec $\sum_{m=1}^{M}\pi_m^s=1$.

Un locuteur est donc modélisé par un ensemble de paramètres noté λ_s [43] :

$$\lambda_s=\left(\pi_m^s,\mu_m^s,\sum_m^s\right) \qquad \text{(I.17)}$$

A.1) Apprentissage du modèle

Il s'agit, lors de la phase d'apprentissage, d'estimer l'ensemble λ des paramètres d'un modèle GMM du locuteur. La méthode conventionnelle est celle du maximum de

vraisemblance (MV) dont le but est de déterminer les paramètres de modèle qui maximisent la vraisemblance des données d'apprentissage. Pour une séquence de N vecteurs d'apprentissage $X = (x_1, ..., x_N)$, la vraisemblance du modèle GMM est [43] :

$$p(X/\lambda) = \prod_{n=1}^{N} p(x_n/\lambda) = \prod_{n=1}^{N} \sum_{m=1}^{M} p(x_n/\pi_m, \mu_m, \Sigma_m) \qquad \text{(I.18)}$$

En remplaçant l'expression de $p(x_n/\lambda)$ on obtient une expression complexe de la vraisemblance et il n'y a malheureusement pas de solution analytique à ce problème [43]. De plus, le calcul de cette expression conduit au logarithme d'une somme et à une fonction non linéaire des paramètres du modèle λ ce qui rend la maximisation directe très difficile [43].

Cependant, le variable indicatrice *m* est une donnée constitutive du problème qui présente l'inconvénient de ne pouvoir être observée en pratique : on observe des réalisations des vecteurs aléatoires x_n sans savoir de manière certaine quelle est la classe du mélange associée à chaque observation. Au sens de l'algorithme EM, le variable *m* Constitue une donnée latente, c'est-à-dire fortement suggéré par le problème considéré. L'introduction de ces données non observées permet de résoudre de manière élégante un problème d'estimation relativement complexe et que ce type de problème est adapté à l'algorithme d'apprentissage EM [43].

Apprentissage par Maximum de vraisemblance (EM)

Le problème de l'algorithme EM (Expectation-Maximisation) peut être considéré comme un cas particulier de gradient [43]. Il fait intervenir à la fois des observations ***X*** et des variables manquantes (l'indice de la gaussienne *m* = 1,……, *M*). Cet algorithme maximise, de façon itérative la fonction de vraisemblance .Cet maximisation n'est pas directe, elle fait intervenir la fonction auxiliaire $Q(\theta, \theta^{(t)})$ qui est définie comme étant l'espérance mathématique du logarithme de vraisemblance jointe (incluant les variables observés et les variables cachées) sur l'ensemble complet des variables d'entraînement, calculée sur base des paramètres courants [43], à savoir :

$$Q(\theta, \theta^{(t)}) = \sum_{m=1}^{M} \sum_{n=1}^{N} p(m/x_n, \theta^{(t)}) \log p(x_n, m/\theta) \qquad \text{(I.19)}$$

Où θ désigne l'ensemble des paramètres à estimer (π_m, μ_m , $\sum_m$) et $\theta^{(t)}$ l'ensemble des paramètres estimés à l'itération *t*. Ce qui donne, après calcul [43] :

$$Q(\theta,\theta^{(t)}) = \sum_{m=1}^{M}\sum_{n=1}^{N}\gamma_{n,m}^{(t)}\left[\log\pi_m - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_m|\right] - \sum_{m=1}^{M}\sum_{n=1}^{N}\gamma_{n,m}^{(t)}\left[\frac{1}{2}(x_n-\mu_m)'\Sigma_m^{-1}(x_n-\mu_m)\right] \quad \text{(I.20)}$$

Où : $\gamma_{n,m}^{(t)}$ est une probabilité a posteriori estimée à l'itération *t*

$$\gamma_{n,m}^{(t)} = \frac{\pi_m^{(t)} p(x_n / \mu_m^{(t)}, \Sigma_m^{(t)})}{\sum_{k=1}^{M}\pi_k^{(t)} p(x_n / \mu_k^{(t)}, \Sigma_k^{(t)})} \quad \text{(I.21)}$$

En supposant que $p(x_n / \theta)$: sont des densités gaussiennes à matrices de covariance diagonales, l'expression de la fonction auxiliaire devient [43]:

$$Q(\theta,\theta^{(t)}) = \sum_{m=1}^{M}\sum_{n=1}^{N}\gamma_{n;m}^{(t)}\log\pi_m - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\gamma_{n;m}^{(t)}\left[Cste + \log\sigma_m^2 + \frac{(x_n-\mu_m)^2}{\sigma_m^2}\right] \quad \text{(I.22)}$$

Où σ_m^2 est un élément diagonal de la matrice de covariance.

Les paramètres sont estimés en annulant la dérivée partielle de la fonction auxiliaire Q par rapport à chacun de ceux-ci. Les cas des poids des composantes de mélange π_m est assez simple puisqu'il s'agit de paramètres scalaires. Ceci dit, il faut tenir compte de la contrainte qui existe sur ces paramètres $\sum_{m=1}^{M}\pi_m = 1$. La maximisation sous contrainte se résout simplement en introduisant un multiplicateur de Lagrange associé à cette contrainte et l'on obtient [43] :

$$\pi_m^{(t+1)} = \frac{1}{N}\sum_{n=1}^{N}\gamma_{n,m}^{(t+1)} \quad \text{(I.23)}$$

En ce qui concerne les vecteurs des moyennes, on montre que les formules de réestimations sont données par [43] :

$$\mu_m^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \, x_n}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{I.24}$$

Et pour les variances [43] :

$$\sigma_m^{2(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t1)} (x_n - \mu_m^{(t)})^2}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{I.25}$$

Apprentissage par Maximum A Posteriori

L'algorithme EM est un des algorithmes les plus importants et les plus puissants en estimation statistique. De plus, il bénéficie d'une preuve de convergence garantissant que l'itération de l'étape d'estimation et de maximisation converge vers un maximum de la fonction de vraisemblance. Cependant, ces limites apparaissent lorsqu'on dispose de peu de données. Donc, il est important d'introduire de l'information a priori. Par conséquent, on ne cherche plus à maximiser la vraisemblance des données mais plutôt la probabilité a posteriori.

Apprentissage incrémental des modèles

L'apprentissage incrémental est un apprentissage bayésien simplifié. Il correspond à un apprentissage MAP (Maximum A Posteriori) avec un choix particulier des paramètres a priori. Les formules de ré-estimation, pour une gaussienne *m*, sont les suivantes [43] :

-Les poids des gaussiennes :

$$\pi_m = \frac{n_m^0 + n_m}{\sum_{k=1}^{M} (n_k^0 + n_k)} \tag{I.26}$$

-Les vecteurs des moyennes :

$$\mu_m = \frac{n_m^0 \overline{X_m^0} + n_m \overline{X_m}}{n_m^0 + n_m} \tag{I.27}$$

- Les variances :

$$\sigma_m^2 = \frac{n_m^0 \overline{X_m^0 X_m^{0'}} + n_m \overline{X_m X_m'}}{n_m^0 + n_m} - \mu_m \mu_m' \qquad \text{(I.28)}$$

Dont, n (respectivement n^0) représente le poids, $\overline{X}$ (respectivement $\overline{X}^0$) le moment d'ordre 1 et $\overline{XX}$ (respectivement $\overline{XX^{0'}}$) le moment d'ordre 2 des données à adapter X (respectivement des données initiales X^0). L'apprentissage incrémental consiste à effectuer quelque itération d'apprentissage sur les données d'adaptation en conservant l'information apportée par les données initiales X^0. Dans le cas où de nombreuses données sont disponibles, l'apprentissage incrémental (ou plus généralement l'estimateur MAP) converge vers les estimateurs du maximum de vraisemblance. Il permet d'obtenir de nouveaux modèles avec peu de données. Ces estimées seront plus fiables que celle obtenues par MV étant donné qu'elles intègrent des connaissances a priori. Cette approche est la plus utilisée en RAL en mode indépendant du texte. La valeur du poids initial n^0 est empirique et comprise entre (8-20) [43].

Dans l'étape d'initialisation Les valeurs initiales d'une densité multi gaussienne peuvent être obtenues par différentes méthodes comme par exemple, la QV (Quantification vectorielle) ou par éclatement de gaussiennes. Cette initialisation est suivie ensuite par un apprentissage EM ou par une adaptation incrémentale. En GMM, le modèle initial correspond au modèle de l'UBM [43]

I.6.2.5 Modèles hybrides

Il ya plusieurs modèles hybride qui combines entre les différents modèles. Parmi les systèmes hybrides GMM-SVM pour la vérification du locuteur [46]. Le modèle hybride HMM-GMM [47] peut être utilisé pour la RAP [47] et RAL [48]. Le modèle DTW-HMM [49] est généralement utilisés aussi par la RAP. Le modèle HMM-ANN (HMM- Artificial Neural Networks) est utilisé pour la RAP.

I.6.3 Normalisation des scores

La variabilité inter-sessions induit dans la phase de test une variabilité des scores de vérification. Cependant si le seuil de décision est fixé empiriquement lors de la phase de

développement, est commun à toutes les conditions de test rencontrées et il est indépendant du locuteur. De ce fait, Il est recommandé d'introduire des techniques de normalisation des scores pour renforcer la robustesse des systèmes de RAL. Ces techniques permettent d'atténuer la variabilité des scores. Elles se basent sur l'analyse des distributions des scores des clients et des imposteurs. Généralement, la normalisation suit la forme suivante [1]:

$$\breve{s} = \frac{s - \mu_I}{\sigma_I} \tag{I.29}$$

Où $\breve{s}$ est le score normalisé, s le score original, et μ_I et σ_I sont respectivement la moyenne et l'écart type des scores imposteurs. Les techniques les plus couramment utilisées sont la zéro normalisation (Z-norm) et le test normalization (T-norm) [50], et se différencient par l'estimation des μ_I et σ_I. Leurs combinaisons, une Z-norm suivie par une T-norm et inversement, sont respectivement appelées la ZT-norm et la TZ-norm [50].

I.6.3.1 Choisit de la norme de normalisation

Bien qu'il y ait plusieurs techniques de normalisation, mais reste la norme T-norme la plus répondu et utilisé dans plusieurs recherches. Il ya encore une autre méthode c'est la méthode adaptative AT-norme inventé par STURIM, Douglas E. et REYNOLDS [51].

Dans notre thèse on a utilisé la norme de T-norme [50].

I.7 Décision et mesure des performances

Bien qui 'il y a deux tâches principales en reconnaissance du locuteur :

1 - Pour l'identification du locuteur on a la relation suivante [1]:

$$I_c = \frac{Nombre\ de\ tests\ correctement\ identifiés}{Nombre\ total\ de\ tentatives} \tag{I.30}$$

et [1] :

$$I_i = \frac{Nombre\ de\ tests\ mal\ identifiés}{Nombre\ total\ de\ tentatives} \tag{I.31}$$

Les performances du système d'identification sont données en termes de taux d'identification correcte I_c ou incorrecte I_i.

2- La vérification du locuteur. C'est une décision en tout ou rien. Les performances de la vérification du locuteur sont données en termes des faux rejets FR et de fausses acceptations FA [1] :

$$FR = \frac{Nombre\ \ de\ \ tentatives\ \ d'abonnésrejetées}{Nombre\ \ total\ \ de\ \ tentatives\ \ d'abonnés} \tag{I.32}$$

$$FA = \frac{Nombre\ \ de\ \ tentatives\ \ d'imposteur\ acceptées}{Nombre\ \ total\ \ de\ \ tentatives\ \ d'imposteurs} \tag{I.33}$$

I.7.1 Distances et mesures de distance

Il est possible d'utiliser toutes les distances classiques, les distances de Minkovski, parmi lesquelles la distance euclidienne, et la distance de Mahalanobis qui normalise les coefficients par leur matrice de covariance. Des distances spécifiques aux espaces de représentation de parole existent aussi, comme les distances cepstrales pondérées, la mesure d'Itakura [52] pour comparer les modèles auto-régressifs, ou encore la distance par appariement de Pics Spectraux [53].

I.8 Conclusion *:*

Dans ce chapitre nous avons décri les différentes classes de paramètres de l'analyse acoustique (les paramètres prosodiques, les paramètres d'analyse spectrale et les paramètres exploitant la dynamique du signal de parole). Le signal de parole est un processus aléatoire non stationnaire à long terme. Un système de reconnaissance automatique du locuteur, quelle que soit la tâche considérée, se résume en trois étapes principales : l'analyse acoustique du signal parole, la modélisation du locuteur et la décision soit une décision et vérification. Également, tout système de RAL dépend de la technique d'extraction de paramètres utilisé, modélisation, décision et ainsi la phase de prétraitement.

Chapitre 2

Réseaux et dégradations

Sommaire

II.1 Introduction

Une des caractéristiques les plus intéressantes des humains est leur capacité de communiquer des idées au moyen de la parole. L'être humain était toujours attiré par la possibilité de créer des machines capables de produire et du fait de reconnaître le locuteur, imitant nous-mêmes. L'objectif final d'un système de RAL est la communication homme-machine. Ce moyen naturel d'interaction a trouvé de nombreuses applications en raison du développement rapide des différents matériels et logiciels. Les plus importants sont l'accès aux systèmes d'information, d'aide aux handicapés, ou le contrôle du système à distance.

Comme mentionné précédemment, un système de RSR (Remote Speaker Recognition) diffère d'un système classique de RAL par leurs mises en œuvre sur les réseaux numériques. À cette fin, ce chapitre se concentre sur les différentes architectures des systèmes de reconnaissance automatique du locuteur à distance à travers les réseaux (mobile, sans-fil, Internet) en tenant compte des dégradations que peuvent subir les canaux de transmissions (phénomène multi trajet, bruits, ...).

II.2 RSR sur les canaux numériques

Il y a plusieurs possibilités pour la mise en œuvre d'un système de RSR sur un canal numérique. Dans la première approche, généralement connue comme la reconnaissance du locuteur/speech dans les réseaux (Network Speaker/speech Recognition NSR) [15], le système de reconnaissance réside dans le réseau de la perspective du client. Dans ce cas, la parole est compressée par un codec (codeur-décodeur) de la parole afin de permettre une transmission de faible débit binaire et/ou d'utiliser un canal du trafic existant des paroles (comme dans le cas de la téléphonie mobile).

La reconnaissance est habituellement réalisée sur les paramètres extraits du signal parole décodé, bien qu'il soit également possible d'extraire les paramètres de reconnaissance directement à partir des paramètres du codec. La figure II.1 montre un schéma de cette architecture du système. Dans le cas où la mise en œuvre est sur un réseau IP, on peut employer un codec VoIP (Voice over IP) [15].

Bien qu'il soit également possible d'extraire les caractéristiques de reconnaissance directement à partir des paramètres du codec, c'est la deuxième approche connue sous le nom de reconnaissance du locuteur/speech distribué (distributed speech/speaker recognition) [15]. Le schéma conceptuel du DSR est illustré à la Figure II.2.

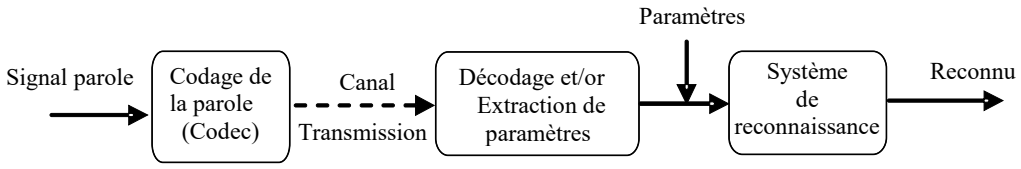


Figure II.1 Schéma d'un système de reconnaissance du locuteur/speech dans le réseau (RSR) (Network Speaker/speech Recognition NSR)

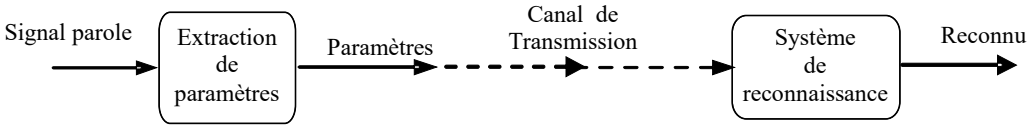


Figure II.2 Schéma d'un système de reconnaissance du locuteur/speech distribué (DSR)

Dans notre travail (Thèse), on a adopté la RAL conformément à la figure II.1. Notre système de reconnaissance du locuteur à travers un canal de communication est décrit en détail dans le chapitre III et VI.

II.3 Les réseaux et dégradations

Les systèmes de RSR dont nous nous occupons exigeant un réseau numérique pour leur déploiement. Généralement, c'est un réseau de téléphonie mobile ou un réseau IP (Internet Protocole). Nous introduisons les traits fondamentaux de ces réseaux qui sont essentiels pour le développement de RSR. On donne un aperçu aux canaux de transmission utilisés par ces réseaux et la dégradation qu'ils habituellement subissent. Les modèles des canaux de transmissions les plus utilisés sont introduits dans les sous-sections qui suivent

II.3.1 Le mobile et le réseau sans fil

La téléphonie mobile numérique a été principalement développé et déployé durant les années quatre-vingt et quatre-vingt-dix, poussés par la forte demande pour cette technique et la nécessité d'un système unifiés pour elle afin de réduire les coûts et d'offrir une qualité maximale de service (QoS) (Quality of Services). En Europe, le mobile GSM (Global System of Mobile communication) a été créé en 1982 par la CEPT (Conférence européenne des postes et télécommunications) pour la standardisation d'un système unifié de radio télécommunication de deuxième génération (2G) et dans la bande de 900 MHz [15, 54]. Versions de GSM ont été également créés dans les bandes de 1800 et 1900 MHz (systèmes DCS1800 et DCS1900). Le travail du groupe GSM a pris une dizaine d'années, période pendant laquelle le GSM acquit le système global de sens pour mobile, et ses

résultats ont été publiés comme un ensemble de normes de l'ETSI (European Telecommunications Standards Institute). Actuellement, cette norme est mise en œuvre dans presque le monde entier [55].

La troisième génération (3G) des communications mobiles est un concept identifié avec l'IMT-2000 (International Mobile Telecommunications, 2000). Il s'agit d'une série de recommandations qui, en fait, a permis d'obtenir plusieurs systèmes 3G comme le CDMA -2000 (CDMA-Code Division Multiple Access) et la téléphonie mobile universelle UMTS (Universal Mobile Telephone System). UMTS a été lancé en 1998 par une association de plusieurs organisations appelées 3GPP.

La vitesse de transmission (de 2G vers 3G) des données s'est augmentée de 144 kbit/s à 2 Mbit/s [15]. L'interface radio UTRA (UMTS Terrestrial Radio Access) utilise deux modes différents pour la liaison radio: TDD (Time Division Duplex- duplex à division de temps) pour les bandes 1900 – 1920 MHz et 2010 – 2025 MHz ainsi pour la liaison radio FDD (Frequency Division Duplex -division de fréquence duplex) pour les bandes 1920-1980 MHz et 2110-2170 MHz. La technique d'accès est CDMA avec deux variantes: W-CDMA (wide band CDMA) et une combinaison TDMA (Time Division Multiple Accès)/CDMA pour TDD.

Il y a en conséquence d'autres générations de réseaux mobiles comme la 4e génération (4 G) [56] et actuellement les recherches tendent à réaliser le 5G [55]. En télécommunications, les 4 G, permet le « très haut débit», c'est-à-dire des transmissions de données à des débits théoriques supérieurs à 100 Mb /s, voire supérieurs à 1 Gb /s (débit minimum défini par l'ITU-T - International Télécommunications Union - Telecoms - Union internationale des télécommunications). En pratique, les débits sont de l'ordre de quelques dizaines de Mb/s selon le nombre d'utilisateurs, puisque la bande passante est partagée entre les terminaux actifs des utilisateurs présentes dans une même cellule radio. Une des particularités de la 4G est d'avoir un « cœur de réseau » basé sur IP et de ne plus offrir de mode commuté (établissement d'un circuit pour transmettre un appel "voix"), ce qui signifie que les communications téléphoniques utiliseront la voix sur IP (en mode paquet).

Les sociétés actuellement ont publié des spécifications qui imposent des téléphones pour pouvoir exploiter son réseau 4G LTE (Long Term Evolution). La $5^{iéme}$ génération de standards pour la téléphonie mobile, faisant suite à la 4G. La technique 5G pourrait permettre des débits de télécommunication mobile, de plusieurs gigabytes de données par seconde, soit jusqu'à 1 000 fois plus rapides que les réseaux mobiles en 2010.

Bien que les systèmes RSR exigent un aperçu sur les réseaux sans fil, on a également mentionné ici deux normes : LAN (local area network-réseau local), IEEE 802.11 (WiFi) et Bluetooth. IEEE 802.11 comprend trois normes : 802. 11 à (54 Mbit/s), 802. 11 b (11 Mbit/s) et 802.11 g (plus de 20 Mbit/s). Par exemple, IEEE 802. 11 b est une technique d'accès de haute puissance qui utilise des séquences à spectre étalé (Direct-Séquence Spread Spectrum (DSSS, semblable au CDMA), à une fréquence de 2,4 GHz et fournit jusqu'à 11 Mbit/s [15]. La figure II.3 donne un schéma général d'un système d'information à travers le réseau (IP, Mobile).

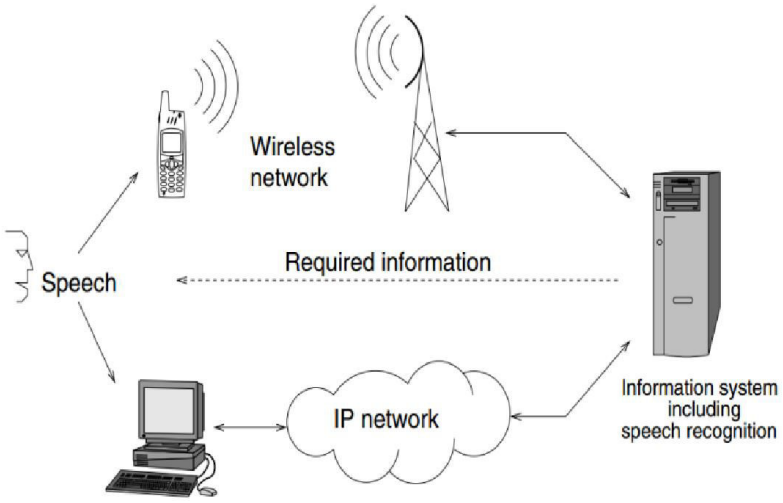


Figure II.3 Schéma général d'un système d'informations parole/locuteur à travers les réseaux (IP, Mobile) [15]

II.3.1.1 Dégradation dans les réseaux sans fil

Lorsque le signal parole est transmis sur un réseau sans fil, il y a deux sources possibles de dégradation qui peuvent dominer sur un système RSR [15] : codage et la transmission de la parole. Nous allons analyser l'effet de codage de la parole dans le chapitre suivant et on se concentrera dans ce chapitre sur le deuxième type de distorsion, introduite par le canal de transmission sans fil.

En premier lieu, le signal transmis est souvent dégradé par un bruit additif inhérent à ce moyen de transmission. Le bruit peut être très destructeur lorsque le récepteur s'éloigne de l'émetteur, puisque l'intensité du signal diminue. Ceci est connu comme perte de chemin d'accès (Path loss). Cependant, la dégradation la plus destructrice du canal radio est due au phénomène de chemins d'accès multiples (Multipath phenomenon) dont la figure II.4 illustre cet aspect. Les ondes radio atteignent l'antenne du récepteur par des chemins différents. Ses différents signaux ont des phases et des amplitudes différentes. Le résultat

est que le signal reçu a un évanouissement (Fading), puisqu'il peut varier son amplitude et la phase d'un endroit à l'autre qui est placée assez proches.

Un évanouissement est essentiellement un phénomène spatial, bien qu'elle se manifeste dans le domaine temporel quand le mobile bouge.

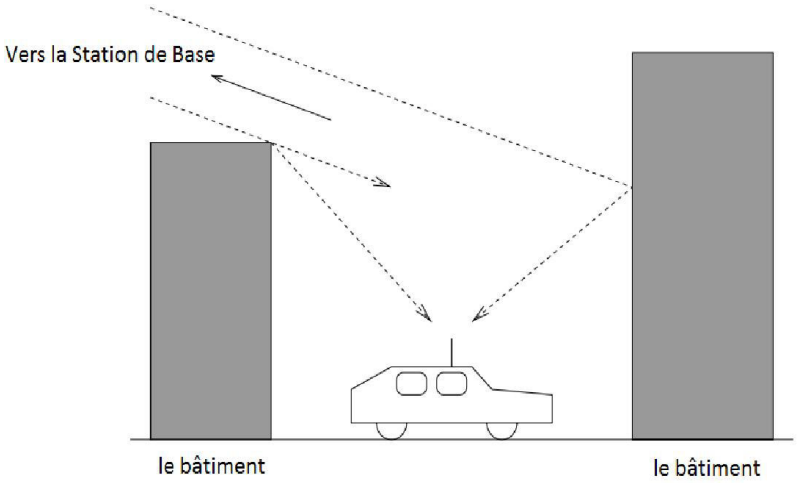


Figure II.4 Illustration du phénomène multi trajet.

Grâce aux trajets multiples, les chemins d'accès différents peuvent contribuer constructivement ou destructivement pour le 1 signal reçu. Ces fluctuations rapides dans l'amplitude et de la phase sont connues comme évanouissement rapide ou évanouissement de Rayleigh. En outre, il peut apparaître comme évanouissement lent (Slow fading) en raison de la perte partielle de trajet produite par de grands obstacles. Évanouissement peut aussi être classé comme plat (flat), quand il affecte également tous les composants de fréquences, et comme sélective (sélective) si elle affecte inégalement ses composantes. Des évanouissements profonds peuvent produire des grandes erreurs, qui pourraient tout à fait dommageable pour une application telle que la reconnaissance du locuteur à travers les canaux de transmission (RSR).

Cependant, avec l'émergence de nouveaux réseaux sans fil utilisant les techniques d'accès cité ci-dessous:

- CDMA ou W-CDMA (Wide band Code division Multiple Accès pour 3G)
- OFDMA (Orthogonal Frequency Division Multiple Access) est une technique de multiplexage et de codage des données utilisées principalement dans les réseaux de téléphonie mobile de 4e génération ou ce codage radio associe les multiplexages en fréquence et temporel ; c'est-à-dire les modes « accès multiple par répartition en fréquence » (AMRF ou en anglais FDMA) et « Accès multiple à répartition dans le temps » (AMRT ou en anglais TDMA).

- SC-FDMA (en anglais «Single-Carrier Frequency Division Multiple Access») est une technique de codage radio numérique utilisée notamment dans les réseaux de téléphonie mobile de 4^{e} génération (LTE); elle utilise simultanément les techniques de multiplexages de type accès multiple par répartition en fréquence et celui par accès multiple à répartition dans le temps (multiplexage fréquentiel et temporel). Le SC-FDMA a attiré l'attention comme une alternative séduisante à l'OFDM et à l'OFDMA, particulièrement dans les communications terre-satellite et dans le sens de transmission montant des réseaux mobiles 4G LTE et LTE-Advanced [57].

En outre, la dégradation due aux interférences d'accès multiple doit également être considérée. Il y a des travaux qui ont étudié ces effets sur le RSR dans [58] et [59]. Erreurs de canal peuvent être prévenues par différentes techniques: la diversité, égalisation adaptative, Codage de canal (Channel coding) et entrelacement.

A) Caractérisation de bruit additif et l'évanouissement (Fading).

Pour une transmission efficace dans un réseau sans fil, les informations sont transmises en utilisant la modulation numérique, La figure II.5 illustre un diagramme général pour la transmission sans fil. Le processus de modulation implique la modification de certains paramètres d'une onde porteuse, obtenant ainsi une série de signaux pour la transmission sans fil. Supposant qu'on veut transmettre un symbole (chaque T [en seconde]) d'un alphabet $\{m_i, i = 1.....M\}$, le cas binaire de (0, 1) est un cas particulier avec $M = 2$.

Un signal $x_i(t)$ de durée T, adapté pour la transmission, est assigné à chaque symbole m_i. La figure II.5 montre le processus de transmission. Après avoir traversé le canal de transmission, nous ne recevons pas le signal original $x_i(t)$, mais une version altérée $y(t)$ est reçu. Et alors, le symbole reçu $\hat{m}$ pourrait différent de m_i si le canal de transmission est suffisamment distordu. Le modèle de canal fréquemment utilisé dans la littérature pour le développement et le test des systèmes de communication est le canal à bruit blanc gaussien (Additive White Gaussian Noise-AWGN) [60, 61]:

$$y(t) = x_i(t) + n(t) \qquad \text{(II.1)}$$

Où $n(t)$ est un bruit blanc gaussien de moyenne égale à zéro et de variance $\sigma_n^2 = N_0/2$ (N_0 est la puissance de bruit), ce bruit est assez fréquent dans les différents systèmes de communication et largement utilisé pour l'analyse du système en raison de sa traçabilité

mathématique. L'évanouissement (le fading) peut être modélisé comme une enveloppe d'un signal aléatoire a et de phase aléatoire θ au signal transmis:

$$\tilde{y}(t) = a \times e^{-j\theta} \times \tilde{x}_i(t) + \tilde{n}(t) \tag{II.2}$$

où le tilde indique l'utilisation d'une notation complexe, $f(t) = \mathrm{Re}\left[\tilde{f}(t)\right]$, un fond de AWGN a également été pris en considération en introduisant $\tilde{n}(t)$. Quand il n'y a aucun élément dominant reçu, l'enveloppe est la distribution de Rayleigh [61]:

$$p(a) = \frac{a}{\sigma^2} \exp\left(\frac{-a^2}{2\sigma^2}\right) \tag{II.3}$$

Dont : $2\sigma^2 = E\left[a^2\right]$ est la puissance moyenne de fading (évanouissement), et la phase a une distribution uniforme. Dans ce cas, nous avons le canal de fading de Rayleigh, qui s'est avéré être proche de la réalité pour le traitement des canaux à évanouissement. Lorsqu'il y a une composante dominante (line-of-sight), l'enveloppe suit une distribution de Ricean [61], et nous obtenons le canal de fading Ricean [61].

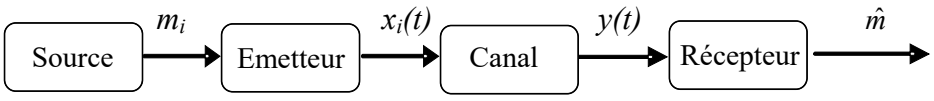


Figure II.5 Diagramme général pour la transmission sans fil.

B) Évaluation de Condition du canal

Une mesure générale de la condition du canal est le rapport signal sur bruit (Signal-to-Noise-Ratio SNR), qui est défini comme le rapport de la puissance moyenne du signal à la puissance moyenne du bruit. Dans le cas où la dégradation causée par des interférences de canaux adjacents ou co-canal, on prend le facteur (C / I) en considération (I: interférences) [15]. Une indication intéressante de performance d'un système de communication numérique est le p_e (la probabilité moyenne d'erreur de symbole). Il est facile de prouver, pour le canal AWGN que [15]:

$$p_e = \frac{1}{2} erfc\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{II.4}$$

Où: *erfc* est la fonction d'erreur [63]:

$$erfc(z) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1) \times n!} \times z^{2n+1} \quad \text{(II.5)}$$

Dans le cas de la modulation par déplacement de phase binaire (Binary phase shift keying- BPSK) on a:

$$p_e = erfc\left(\sqrt{\frac{E_b}{N_0}}\right) \quad \text{(II.6)}$$

Dans le cas de la modulation Quadrature par déplacement de phase (Quadrature Phase Shift Keying-QPSK). E_b : est l'énergie pour un bit transmit: $E_b = E$ pour BPSK, et $E_b = E/2$ pour QPSK, (chaque symbole a deux bits).

L'équation II.4 est une approximation de l'erreur pour le cas de la modulation GSM GMSK (Gaussian minimum-shift keying).

Le rapport E_b/N_0 dépend de SNR du canal de communication [62]:

$$SNR = \frac{E_b}{N_0} \frac{R_s}{B} \quad \text{(II.7)}$$

où: Rs est le débit binaire qui est une mesure de la quantité de données numériques transmises par unité de temps. Il est le plus souvent exprimé en bits par seconde (bit/s, b/s ou bps). B est la bande passante. Un autre facteur de mérite est le taux d'erreur binaire (Bit Erreur Rate - BER) qui, en général, diffère de p_e depuis chaque symbole peut contenir plusieurs bits. Ainsi, BER et p_e coïncident dans le cas BPSK, tandis que pour une modulation QPSK, il peut être obtenu par [62]:

$$BER = \frac{1}{2} erfc\left(\sqrt{\frac{E_b}{N_0}}\right) \quad \text{(II.8)}$$

Nous voyons que QPSK transmet deux fois plus d'informations que BPSK pour le même BER et par conséquent, pour le même rapport du E_b/N_0. Ainsi, dans ce sens, QPSK

a une meilleure performance que BPSK [15]. Dans le cas d'un canal de Rayleigh et BPSK, on peut montrer que, si l'évanouissement est suffisamment lent pour que la phase puisse être déterminée avec précision, la probabilité d'erreur est [15]:

$$p_e = \frac{1}{2}\left(1 - \sqrt{\frac{\gamma_0}{1+\gamma_0}}\right) \tag{II.9}$$

où: $$\gamma_0 = E\left[a^2\right] E_b / N_0 \tag{II.10}$$

Dans le cas de BPSK sur un canal fading, il peut être démontré que la probabilité de d'erreur $p_e(y)$ d'un signal reçu y peut être calculée comme [15]:

$$p_e(y) = \frac{1}{1+\exp|L_c y|} \tag{II.11}$$

Avec [15]:

$$L_c = 4a\frac{E_b}{N_0} \tag{II.12}$$

Pour un facteur d'évanouissement est supposé connaître au récepteur (a = 1 correspond au canal AWGN). $|L_c y|$ peut être considéré comme une mesure de la fiabilité du bit reçu.

II.3.1.2 RSR sur les réseaux mobiles

Le concept de RSR est développé en parallèle avec le progrès des réseaux mobiles à commutation de circuits (circuit-switched) et en utilisant une architecture de NSR, (travaux d'EULER, S. 1994 [64] pour le cas de reconnaissance de la parole et Besacier, Laurent [65] pour le cas de RAL), cela en utilisant les codecs de réseau et le codage de canal. Les performances de cette architecture (NSR) peuvent être très vulnérables au codage de la parole et dégradations de canal [15].

DSR est apparu comme une alternative pour pallier ces dégradations sur les canaux à commutation de circuits. Cependant, avec l'évolution des réseaux mobiles vers la 3G, 4G et 5G, les réseaux mobiles à commutation de paquets (tels que GPRS) semblent plus appropriés pour l'intégration de la RSR. Plus de détails sur les principes des systèmes de transmission de signal parole à travers le mobile sont dans (VARY, Peter. 2006 [66]). La mise en œuvre d'un system de RSR sur le réseau mobile est traitée dans le chapitre VI.

II.3.2 Le réseau IP

Il s'agit de la transmission de la voix à travers le réseau IP (Voix over IP - VoIP). La téléphonie sur IP exploite un réseau de données IP pour offrir des communications vocales à l'ensemble de l'entreprise sur un réseau unique voix et données. Cette convergence des services de communication données, voix, et vidéo sur un réseau unique, s'accompagne des avantages liés à la réduction des coûts d'investissement, à la simplification des procédures d'assistance et de configuration…..

L'origine de l'Internet et IP est le réseau ARPANET (Advanced Research Projects Agency Network » qui est le premier réseau à transfert de paquets développé aux États-Unis par la DARPA (Defense Advanced Research Projects Agency) de commutation de paquets, développée au cours des années soixante-dix. La commutation de paquets se pose comme une alternative au circuit traditionnel à commutation permettre le partage des ressources entre les ordinateurs. L'unité de transmission dans un réseau à commutation de paquets est un bloc de données appelé paquet. Un message peut être transmis dans un paquet unique ou divisé en plusieurs paquets, qui sont transmis de façon indépendante. Bien que les paquets puissent suivre un itinéraire préplanifié de leur source (approche de circuit virtuel), nous allons seulement faire attention à l'approche du datagramme, pour lesquels les différents paquets d'un message donné peuvent suivre des routes différentes dans le réseau jusqu'à leur destination. Dans ce cas, chaque paquet se compose d'un en-tête (contenant l'adresse de destination) et les données utiles à transmettre. Contrairement à la commutation de circuits, il n'y a pas un itinéraire préétabli entre l'émetteur et le récepteur. Par conséquent, les paquets peuvent être transmis sans attendre une connexion. D'autre part, il n'est pas possible de s'assurer si un paquet donné va atteindre sa destination (ou s'il va arriver). ARPANET utilise commutateurs qui ont été connectés au minimum à deux ordinateurs. Ainsi, les paquets avaient différents itinéraires pour atteindre la destination. Chaque commutateur avait une table de routage, en précisant de quelle manière un paquet devrait suivre, et stocker en mémoire les paquets entrants jusqu'à ce qu'ils aient été retransmis. Après le succès de ARPANET, la nécessité se pose pour la connexion de différents réseaux informatiques, pour laquelle il était nécessaire d'élaborer des protocoles permis la compatibilité. La connexion est possible grâce à un dispositif de routage qui s'adapte le format des informations en plus de contenir l'information de routage nécessaire pour les deux réseaux. Ce problème n'est apparemment simple, car il s'avère de plus en plus complexe lorsque le nombre de réseaux interconnectés se développe. Le protocole de contrôle de transmission (TCP) / IP a résolu le problème, de sorte que l'échange

d'informations entre les réseaux est transparent et l'utilisateur peut les voir comme un seul réseau virtuel. TCP / IP utilise la même idée de commutation de paquets ARPANET, par conséquent, les dispositifs de routage n'ont pas besoin de stocker des informations sur les états des utilisateurs ou les flux d'informations. Les spécifications du protocole TCP / IP ont été développés par l'Internet Engineering Task Force (Internet Engineering Task Force-IETF) et sont connues comme RFC (Request For Comment - demande de commentaire).

Pour ce qui est RSR, le réseau IP apparaît comme une plateforme naturelle pour sa mise en œuvre, en raison de leur/ architecture client-serveur, qui est identique à celle de la RSR (voir la figure II.3). Dans ce sens-là l'établissement d'un système de reconnaissance du locuteur à travers ce réseau est très intéressant. il y a plusieurs tentatives d'implémentations des systèmes de reconnaissance (parole /speaker) à traves IP (voir [67, 68, 69, 70, 71]), mais la fiabilité de ces systèmes dépend des codecs utilisés (où il y a plusieurs standards discutés dans le chapitre III) : les techniques d'extraction de paramètres, prétraitement (front-end-processing) et les techniques de dé-bruitage. Dans le chapitre IV, on a proposé une architecture pour la transmission de la voix à travers IP avec une comparaison des techniques de codages (suivant la standardisation de l'IUT), et une technique de détection parole/non-parole [72]

II.3.2.1 Le protocole TCP/IP

TCP/IP implique une famille de protocoles conçus pour effectuer différentes tâches et fournir des services [15], le tableau II.1 montre cette structure de couche et de ses protocoles. Le tableau montre comment les différents protocoles TCP / IP sont appliqués l'une sur l'autre. Les couches TCP / IP reposent sur une couche physique (signaux précisant, débit de données), qui sont également représentés sur le tableau.

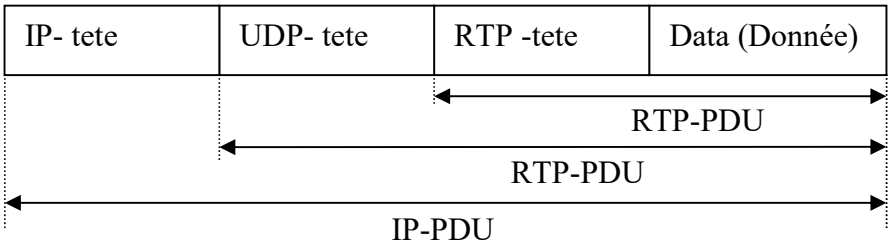
Contrairement aux réseaux de commutation de paquets, qui sont utilisés pour la transmission de données et qui n'ont pas d'exigences temporelles particulières, les réseaux à commutation de circuits ont été traditionnellement utilisés pour les données ayant des exigences en temps réel. Cependant, les améliorations de la vitesse de traitement informatique, le développement de puissantes techniques de compression de données et l'augmentation de la bande passante disponible sont des faits qui permettent la mise en œuvre d'applications en temps réel sur des réseaux IP tels que, par exemple, la téléphonie IP (la voix sur IP (Vo Ip) et RSR. Il y a un protocole spécial pour la transmission de

données en temps réel appelé protocole en temps réel (RTP). Comme le montre le tableau II.1, il s'appuie sur le protocole de datagramme utilisateur (UDP), et sur le protocole IP. Un paquet RTP et sa structure sont représentés sur la figure II.6 il contient trois PDU intégrés (unité des protocoles de données - Protocole Data Unit) [15].

Tableau II.1 Représente la structure de couche TCP/IP et les protocoles communs.

Couches	Protocoles	
Applications	FTP, HTTP, Telnet, VoIP,...	
Transports	TCP	RTP
		UDP
Interconnexion	IP	
Interface de réseaux	Ethernet, Wifi, ATM,....	
Physique	Les caractéristiques de transmission	

Figure II.6 Format de paquets en utilisant RTP [15]



a) Les fonctions de RTP

RTP, est un protocole adapté aux applications présentant des propriétés temps réel. Il permet ainsi de [72]:

- reconstituer la base de temps des flux.
- mettre en place un séquencement des paquets par une numérotation et ceci pour permettre la détection des paquets perdus.
- identifier le contenu des données pour leur associer un transport sécurisé.
- l'identification de la source c'est à dire l'identification de l'expéditeur du paquet. (dans un multicast l'identité de la source doit être connue et déterminée).
- transporter les applications audio et vidéo dans des trames (avec des dimensions dépendantes des codecs qui effectuent la numérisation). Ces trames sont incluses dans des paquets afin d'être transportées et doivent de ce fait être récupérées

facilement au moment de la phase de dépaquétisation afin que l'application soit décodée correctement.

En revanche, ce n'est pas "la solution" qui permettrait d'obtenir des transmissions temps réel sur IP. En effet, il ne procure pas de : Réservation de ressources sur le réseau et fiabilité des échanges (pas de retransmission automatique et de régulation du débit).

II.3.2.2 Dégradation dans les réseaux IP

Contrairement aux réseaux sans fil, des erreurs de bit dans les réseaux IP peuvent être négligées pour un certain nombre de raisons. La charge utile peut contenir la détection d'erreurs et / ou des mécanismes de correction. De plus, l'en-tête UDP peut contenir un mécanisme de vérification des erreurs pour les données utiles. Enfin, il faut tenir compte du fait que les réseaux sous-jacents sont généralement assez fiables [15]. Le réseau typique a des liens de câbles à haute valeurs SNR pour le canal. On peut considérer un réseau IP comme un support de transmission où la dégradation peut apparaître en raison des inconvénients inévitables à sa structure de commutation de paquets, qui est principalement de latence, la propagation de temps et la perte des paquets. Les deux premiers peuvent être traités au moyen de l'introduction de retard de décodage, ce qui peut les absorber. Paquets qui se larguent sont considérés comme perdus [15]. Délai et la propagation de temps sont importants dans des applications telles que le VoIP, où il est prévu de maintenir l'interactivité d'une conversation. Pour une application RSR, une réponse immédiate à partir du serveur, bien que non indispensable, est également souhaitable. Par conséquent, les pertes de paquets apparaissent comme la principale source de dégradation qui peut affecter les performances de la reconnaissance [71].

En plus de paquets en retard, les pertes de paquets peuvent avoir lieu au niveau des dispositifs de routeur. Figure II.7 montre comment ce dispositif fonctionne: il a plusieurs files d'attente d'entrée et de sortie, un commutateur de paquet (packet-switcher) et une logique de commutation (a switching logic) qui décide (en utilisant l'adresse IP) dans lequel dans la file d'attente de sortie, un paquet donné doit être placé. La fonction des files d'attente est d'adapter les différentes vitesses de transmission des réseaux interconnectés. Ils peuvent suivre des politiques simples comme premier arrivé, premier servi (PAPS, first-come-first-serve - FCFS) ou d'autres plus sophistiqués tels que les files d'attente équitables pondérées (weighted fair queuing-WFQ), ce qui implique un partage équitable de la sortie. Pertes de paquets peuvent apparaître à cause de plusieurs circonstances comme mentionnées dans (Kurose et Ross, 2005 [73]):

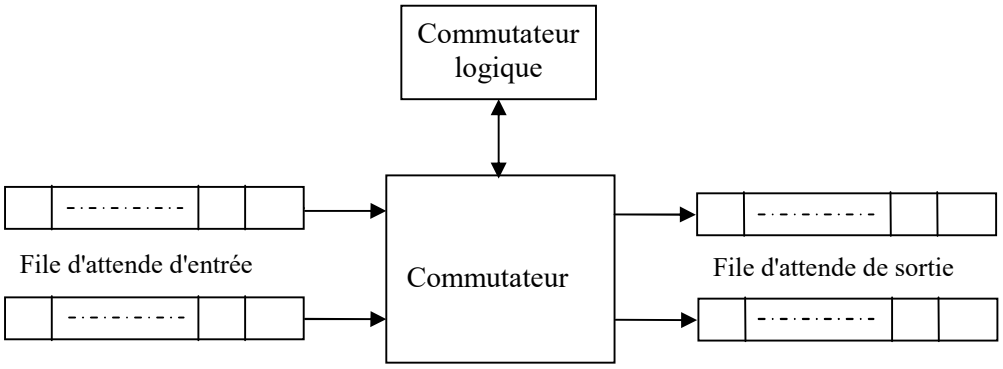


Figure II.7 Schéma d'un dispositif routeur.

II.3.2.3 Modèles de canal à paquets perdues

Il ya plusieurs techniques de modélisation de perte de paquets parmi les: Bernoulli Mode, Gilbert –Elliot Models, Autres Première commande Modèles de Markov (Other First-order Markov Models), Higher-order Markov Models [15].

II.4 Environnement acoustique

Dans cette section on considère la dégradation introduite par l'environnement acoustique c'est-à-dire les dégradations dues à la prise du son (bruits ambiants et variations du canal acoustique). Les systèmes de RAL, en particulier les systèmes de RSR formés à l'aide des signaux parole, peuvent se dégrader de manière significative lorsqu'ils sont utilisés dans des conditions réelles. Il y a plusieurs raisons pour lesquelles la parole dans un environnement réel peut différer de parole propre. L'environnement sonore peut être défini comme l'ensemble des transformations qui modifient le signal parole à partir du moment où il quitte la bouche jusqu'à ce qu'il soit enregistré sous forme numérique. Signal enregistré dans différents environnements acoustiques a des caractéristiques différentes. Le décalage introduit par des variations de l'environnement acoustique est la principale source de la dégradation des systèmes de RAL. Dans cette section, nous nous concentrons sur les deux principales sources de distorsions: le bruit additif et distorsion de canal.

II.4.1 Bruit additive

Le bruit d'additif de terme est tout signal indésirable qui est ajouté au signal désiré. La source la plus commune de bruit est le bruit de fond. C'est communément renvoyé comme le bruit acoustique et est provoqué par: les climatiseurs, les ventilateurs informatiques, les voitures mobiles, murmure confus, environnement train, l'aéroport.... . Un signal capturé

avec un microphone a peu de bruit de fond. Cependant, si un microphone lointain est utilisé ce qui va engendrer une grande quantité de bruit de fond qui peut être enregistrée avec le signal de parole. D'autres types de bruit additif sont comme suit :

1) bruit électromagnétique : provoqué par les appareils électriques.
2) bruit électrostatique : produit par la présence d'un voltage. Les lumières fluorescentes sont la source principale de ce type de bruit.
3) traitement du bruit : en provenant du traitement analogique/numérique du signal parole. Un exemple commun est l'erreur quantification dans le codage numérique.

II.4.1.1 Bruit Blanc Gaussien

Le bruit blanc de terme se réfère à un signal $x(t)$ avec une densité spectrale de puissance plate $S_{xx}(f)$. Cela signifie que $x(t)$ un signal non corrélé [15]:

$$R_{xx}(\tau) = E\left[x(t)x(t+\lambda)\right] = \sigma^2\delta(\tau) \tag{II.12}$$

Dont il est obtenu [15]:

$$S_{xx}(f) = \int_{-\infty}^{\infty} R_{xx}(t)e^{-j2\pi ft}dt = \int_{-\infty}^{\infty} \sigma^2\delta(\tau)e^{-j2\pi ft}dt = \sigma^2 \tag{II.13}$$

Soit σ^2 la variance du processus de bruit. Ce type de bruit peut être généré par une distribution $p(x)$ donnée à partir des échantillons, nous pouvons avoir différents types de bruit blanc selon la distribution sélectionnée. Lorsque $p(x)$ est uniforme, le bruit est un bruit blanc uniforme. Bruit blanc gaussien est obtenu en utilisant une distribution de probabilité gaussienne.

II.4.1.2 Bruit coloré

Bien que le bruit soit un signal aléatoire, il possède des propriétés statistiques caractéristiques. La densité spectrale de puissance en est une, et peut être utilisée pour distinguer les différents types de bruit. Cette classification par la densité spectrale donne une terminologie de « couleurs ». Chaque type est défini par une couleur. Ces définitions sont, en principe, communes aux différentes disciplines pour lesquelles le bruit est un facteur important (comme l'acoustique, l'électrotechnique et la physique) [74].

La plupart des définitions de bruits colorés font état d'un signal présent à toutes les fréquences, et qui possèdent une densité spectrale par unité de largeur de bande (bande passante) proportionnelle à : $\frac{1}{f^{\beta}}$ (où f est la fréquence et β un nombre) [74]. Le bruit blanc est monotone avec β = 0, le bruit rose correspond à β = 1, et le bruit marron à β = 2. dans la littérature on trouve plusieurs modèles de bruit coloré comme, Bruit rouge ou brownien (dans les domaines qui utilisent des définitions précises, la terminologie « bruit rouge », « bruit brownien » ou « bruit brun » fait référence au son ayant une puissance sonore qui décroît de 6 dB par octave lorsque la fréquence augmente), Bruit bleu (La puissance sonore du bruit bleu augmente de 3 db par octave lorsque la fréquence augmente); Bruit Violet (La puissance sonore du bruit violet augmente de 6 dB par octave lorsque la fréquence augmente), Bruit gris (Le bruit gris est un bruit rose soumis à une courbe psycho-acoustique d'intensité constante) [74]. Figure II.8 illustre différent nature de bruit (Rose, Rouge, Blue et Violet).

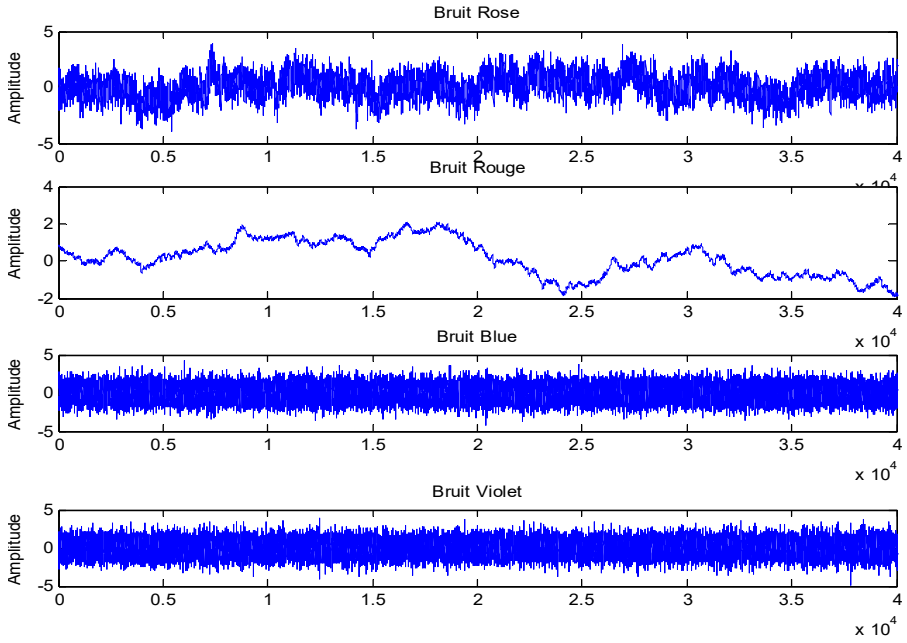


Figure II.8 Différentes natures de bruits (rose, rouge, bleu et violet).

II.4.1.3 Bruits stationnaires et non stationnaires

Un bruit stationnaire est un signal dont les caractéristiques sont constantes au cours du temps [75]. Si les caractéristiques du bruit varient au fil du temps, le bruit est appelé non stationnaire. Au sens strict, il n'y a aucun bruit stationnaire parfait, mais certains bruits sont près stationnaires comme le bruit aérodynamique produit dans les voitures ou les bruits provenant des ventilateurs ou des climatiseurs [75].

II.4.2 Distorsion de Canal

Distorsion du canal est la deuxième source de distorsion acoustique et causée par un changement dans la forme spectrale du signal parole en raison de la réponse en fréquence de la chaîne de transmission acoustique.

II.4.2.1 Réverbération

À moins que le microphone et le haut-parleur soient trouvés dans l'espace libre ou dans une chambre sourde, le microphone capte, ainsi que le signal de la trajectoire acoustique directe, signaux reflètent dans les obstacles à proximité (murs ou autres objets dans la salle) [76]. Un exemple de cette situation est représenté dans la figure II.9. Le signal reçu au niveau du microphone est la somme des signaux reçus par le biais de la voie directe et tous les chemins indirects. Nous devons désigner le signal voyageant le long de la voie directe de d_0 longueur comme $x_0(t)$. Ce signal est reçu avec atténuation inversement proportionnelle à la distance et avec un délai donné par : $\lambda_0 = {d_0}/{c}$, (c : étant la vitesse du son). Le signal reçu par une voie indirecte $x_k(t)$ est reçu avec un retard de la râpe et une amplitude plus faible en raison de la plus longue du chemin. Dans ce cas, l'atténuation est en raison non seulement propagation, mais aussi l'absorption partielle qui se produit à chaque réflexion. Compte tenu de l'effet combiné de l'atténuation et le retard, on écrit [15] :

$$x_0(t) = \left({A}/{d_0}\right) x(t-\tau_0) \tag{II.14}$$

$$x_k(t) = r_k \left({A}/{d_k}\right) x(t-\tau_k) \tag{II.15}$$

r_k étant l'atténuation totale en raison des réflexions sur le chemin de k^{ieme}. Le signal reçu est une combinaison des versions atténuées et retardées du signal original [15]:

$$x(t) = x_0(t) + \sum_{k=1}^{\infty} x_k(t) = \left({A}/{d_0}\right) x(t-\tau_0) + \sum_{k=1}^{\infty} r_k \left({A}/{d_k}\right) x(t-\tau_k) \tag{II.16}$$

En ce qui concerne, le signal reçu sur la voie directe $x_0(t)$, on [15]:

$$x(t) = x_0(t) + \sum_{k=1}^{\infty} r_k \left(d_0 / d_k \right) x_0\left(t - (\tau_k - \tau_0)\right) = h(t) * x_0(t) \qquad \text{(II.17)}$$

La réponse impulsionnelle de la réverbération du système $h(t)$ est [15]:

$$h(t) = 1 + \sum_{k=1}^{\infty} r_k \left(d_0 / d_k \right) \delta\left(t - (\tau_k - \tau_0)\right) \qquad \text{(II.18)}$$

Et la réponse en fréquence du système est [15]: Microphone

$$H(f) = 1 + \int_{-\infty}^{\infty} h(t) \cdot e^{-j 2\pi f t} dt = 1 + \sum_{k=1}^{\infty} r_k \left(d_0 / d_k \right) \cdot e^{-j 2\pi f (\tau_k - \tau_0)} \qquad \text{(II.19)}$$

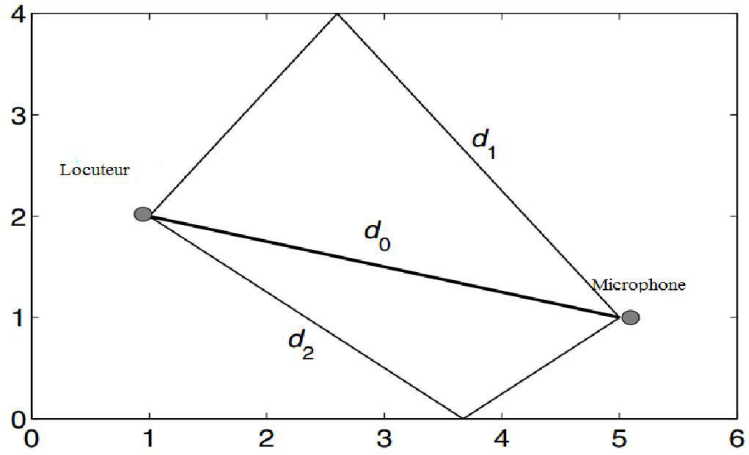


Figure II.9 Situation schématique causant la réverbération en chambre mentionnant le chemin direct d_0, deux chemins indirects d_1 et d_2.

Une étude des méthodes d'élimination du bruit causé par la réverbération est nécessaire pour améliorer le taux de reconnaissance. Dans le chapitre VI on explore différentes techniques d'élimination de bruit dû aux réverbérations.

II.4.3 Modèle de l'environnement acoustique

Comme décrit dans les paragraphes précédents, l'effet de l'environnement sur le signal de parole peut être décrit par deux principales contributions : le bruit additif et la distorsion du canal. Par conséquent, nous pouvons établir le modèle suivant la déformation subie par le signal de parole en raison de l'environnement :

$$y\left(m\right)=x\left(m\right)*h\left(m\right)+n\left(m\right) \tag{II.20}$$

Où $y\left(m\right)$ exprime l'échantillon déformé, $h\left(m\right)$ est la réponse impulsionnelle du canal et $n\left(m\right)$ est le bruit additif. La figure II.10 représente le modèle de l'environnement acoustique.

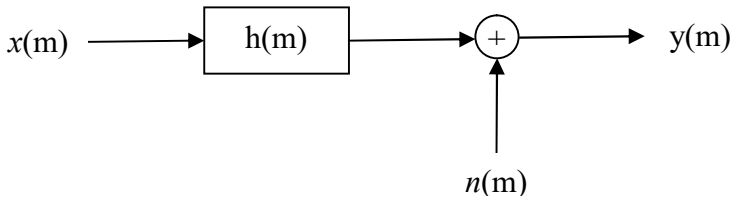


Figure II.10 Modèle de l'environnement acoustique.

Dans le domaine temporel, le signal de parole déformée est obtenu par le produit de convolution du signal original avec la réponse impulsionnelle du canal plus un terme de bruit additif. Dans le domaine spectrale l'équation II.20 devient [15]:

$$\left|Y\left(K\right)\right|^2=\left|X\left(k\right)\right|^2\left|H\left(k\right)\right|^2+\left|N\left(k\right)\right|^2+2\,\mathrm{Re}\left\{X\left(k\right)H\left(k\right)N^*\left(k\right)\right\} \tag{II.21}$$

Le dernier terme dans cette équation est petit et a une valeur attendue de zéro. Si nous utilisons une approche de banc de filtre (filterbanc) pour l'analyse du signal parole, la valeur faite en moyenne du trans-produit est petite et peut être négligée. Donc, l'énergie de production de banc de filtre est [15]:

$$\left|Y_b\right|^2\approx\left|X_b\right|^2\left|H_b\right|^2+\left|N_b\right|^2 \tag{II.22}$$

Dans cette approche, nous avons fait aussi l'hypothèse implicite que la réponse d'impulsion du canal est plus courte que la longueur de fenêtre utilisée pour l'estimation des spectres. Les paramètres les plus fréquemment utilisés dans les systèmes de reconnaissance du locuteur sont basés sur le MFCC, qui est défini comme la transformation discrète de cosinus (DCT- Discret Cosinus transform) des log-énergie à la sortie de l'échèle de mel de banc de filtres. Pour appliquer le modèle d'environnement d'équation (II.20) à ces traits (MFCC), nous définissons d'abord les vecteurs suivants [15]:

$$x=\left[\log\left|X_1\right|^2\log\left|X_2\right|^2.....\log\left|X_M\right|^2\right] \tag{II.23}$$

$$y = \left[\log|Y_1|^2 \log|Y_2|^2 .....\log|Y_M|^2 \right] \quad \text{(II.24)}$$

$$h = \left[\log|H_1|^2 \log|H_2|^2 .....\log|H_M|^2 \right] \quad \text{(II.25)}$$

$$n = \left[\log|N_1|^2 \log|N_2|^2 .....\log|N_M|^2 \right] \quad \text{(II.26)}$$

Où M est le nombre de filtres dans le banc de filtre (c'est le nombre de filtres triangulaire, souvent choisi égale à 24 [77]). Ainsi, les filtres de mel souvent espacés dans la gamme de 0-8000 Hz [77]. Les équations de log-énergie à la sortie du banc de filtre sont [15]:

$$y = \log\left(\exp\left(x + h\right) + \exp(n)\right) = x + h + \log\left(1 + \exp\left(n - x - h\right)\right) \quad \text{(II.27)}$$

Le signal parole bruité y est obtenu comme une combinaison non linéaire du signal parole original x (sans bruit), le bruit ‘n’ et le canal ‘h’. Cette expression peut être utilisée pour prédire l'effet de milieu au signal parole.

L'effet principal de la combinaison non linéaire de bruit sur log-Energie du banc de filtres est la réduction de gamme et un changement de la valeur moyenne [15].

Les coefficients MFCC sont obtenus comme les combinaisons linéaires de log-énergies donc ils subissent aussi une distorsion non linéaire. Les effets principaux sont de nouveau une réduction de gamme et un changement des valeurs moyennes [15].

II.5 Robustesse contre les erreurs de canal de transmission

Les canaux numériques peuvent introduire plusieurs types de dégradation tels que la perte de paquets ou le fait d'évanouissement. Nous avons étudié aussi différents modèles de canal qui peuvent être utiles pour comprendre et simuler l'effet de différents canaux sur les informations transmises. .

Un système de RSR robuste doit être conçu pour qu'il puisse maintenir une performance acceptable en présence de la dégradation du canal. Il est donc nécessaire de développer un ensemble de techniques pour prévenir, corriger et atténuer ses effets. En général, nous les appellerons comme les techniques de récupération. Il peut être possible de classifier et étudier les différentes techniques de récupération selon le cadre de transmission spécifique, c'est-à-dire l'architecture du système choisie, le type de canal considéré

(physique - ou une plus haute couche) et le réseau radio ou fil métallique, circuit - ou circuit du paquet (packet-switched)) sur lequel le système est exécuté. Cependant, les techniques habituelles utilisées sont communément partagées par les différents cadres de transmission, donc il est plus approprié de les classifier en considérant exclusivement le type de techniques elles-mêmes et plus tard spécifier comment ils sont appliqués dans un cadre donné. Une telle classification peut être établie comme suit:

1- **Techniques produites par l'émetteur** : il y a un certain nombre de techniques de codage de canal qui sont convenables pour RSR, comme:

a- le correction d'erreurs avancée (Forward error correction-FEC): ce codeur de canal introduit une redondance contrôlée dans le message issu du codeur de source. Cette redondance connue aussi au niveau du récepteur, permettra la détection et/ou la correction des erreurs dues au bruit inévitable.

b- Entrelacement: les informations à transmettre ont été réordonnés avant la transmission. Ainsi, les erreurs de canal (spécialement les grouper dans les bursts (paquet d'informations)) sont au hasard c'est-à-dire que les erreurs dues au canal n'influent pas fatalement sur un seul burst.

2- **Techniques utilisées par le récepteur:** Nous allons également faire référence à ces techniques que la dissimulation d'erreurs (Error Concealment- CE [66]) ou des techniques d'atténuation. Dans ce cas, le récepteur doit traiter les erreurs du canal sans aucune participation de l'émetteur. Ces techniques sont utiles lorsque les techniques de l'émetteur ne conduisent pas à corriger les données erronées ou perdues ou lorsque l'expéditeur n'est pas en mesure de participer au processus de récupération. Nous allons examiner les trois groupes de techniques: Interpolation, estimation et techniques de reconnaissance de base. Estimation et interpolation sont des techniques de reconstruction, car ils fournissent un remplacement des données endommagées ou perdues. Ces remplacements sont entrés dans la RAL comme s'ils étaient totalement fiables [66].

Les différentes techniques de récupération ne sont généralement pas exclusives, de sorte qu'ils peuvent être combinés pour augmenter la robustesse. En particulier, les informations obtenues lors de décodages du canal peuvent être utilisées par les modules de la CE pour fournir une meilleure performance. En plus la dégradation du canal peut également être traitée par des techniques de corrections. La figure II.11 montre un schéma synoptique de l'ensemble du système de transmission de reconnaissance automatique du locuteur à

distance (RSR) considéré dans ce travail avec plus de détails dans le chapitre VI. Il comprend à la fois le codage de canal et de la CE [66]. La première étape du processus d'encodage est appelé codeur de source et contient les algorithmes de compression (voir chapitre III). Les informations codées dépendent de l'architecture RSR, les caractéristiques vocales dans le cas de DSR et le signal de parole lui-même dans le cas de NSR. Le flux binaire "d" produit par le codeur de source est soumis au codeur de canal, qui introduit les redondances de protection contre les erreurs et génère "c"

Les bits transmis (bit stream) peuvent être dégradés par le canal, de sorte que nous recevons "$\hat{c}$". Le récepteur effectue le décodage du canal et CE [66] où après le décodage du canal on reçoit un flux binaire $\hat{d}$. La CE peut être effectuée par le décodeur de source, par le dispositif de reconnaissance ou par les deux [15].

Une question importante concernant les techniques de récupération est le temps de latence que nous pouvons permettre. Dans un système RSR, une réponse immédiate du système est souhaitable, mais pas indispensable, de façon un petit retard peut être accepté que s'il favorise une meilleure performance de reconnaissance. C'est une différence importante par rapport à d'autres applications telles que la téléphonie mobile ou VoIP, où la latence doit être prise en compte par le concepteur du système.

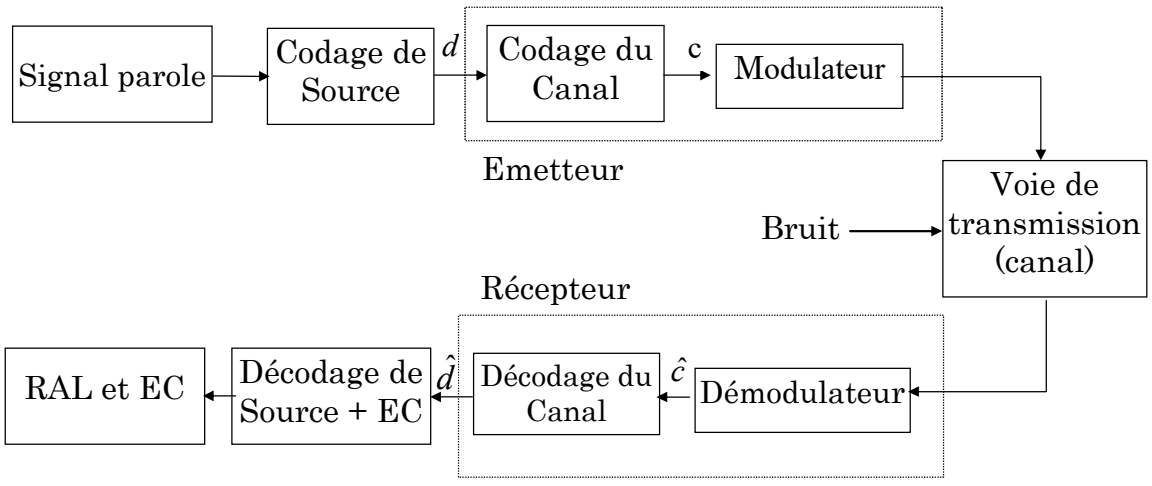


Figure II.11 Schema general de tout système de transmission numérique destine au RSR.

II.5.1 Techniques de codage de canal

Dans cette section, nous allons voir les principaux types de codage de canal, codages FEC et l'entrelacement, appliquées aux systèmes RSR. Les principaux codages de canal

sont les principaux FEC : codes linéaires de blocs, codes cycliques et les codes de convolution [66].

II.5.1.1 Détection des erreurs

Une unité de la parole touchée par un canal de transmission sans fil dégradé peut-être soit erronée soit perdue (transmission par paquets). Bien que les deux types de dégradation fassent réduire les performances de reconnaissance d'un système RSR, l'effet d'une perte est bien plus petit que celui d'une erreur de voie (Bernard et Alwan, 2002 [78]). Une unité perdue peut être facilement détecté (Juste en vérifiant le numéro de commande de paquets) (par exemple en reproduisant le dernier reçu) dans les systèmes DSR, de sorte que la précision de reconnaissance peut être maintenue à un niveau acceptable [15]. Une solution efficace pour les unités erronées est l'utilisation de FEC à fin de les détecter et considérer les unités affectées comme pertes du canal. Quand le décodeur canal exclusivement effectue une détection de l'erreur, la responsabilité de fournir des données adéquates pour remplacer les données erronées ou perdues, tombe sur l'algorithme d'EC qui doit générer ces remplacements comme si les données source ont été effacées par le canal. Dans ces cas, nous allons utiliser le terme de canal d'effacement (D. Proietti, 2002 [79]).

II.5 Conclusion

Ce chapitre a présenté une vue d'ensemble sur la reconnaissance du locuteur sur les réseaux mobile, sans fil et internet. On a donné un aperçu sur le signal parole dans les réseaux (Mobile et IP), dont les performances d'un système de reconnaissance automatique du locuteur à distance se dégradent à cause de la distorsion du canal et la présence de bruits. Pour remédier au maximum à ce problème de dégradation, il est nécessaire d'utiliser un code correcteur (FEC). Un code correcteur est une technique de codage basée sur la redondance destiné à corriger les erreurs de transmission d'une information sur une voie de communication peu fiable. En d'autres termes, le codage ou la compression du signal parole est considéré comme un facteur important dans la dégradation d'information transmise à travers le canal. Bien que le type de codage ou compression soit un facteur important, dans le chapitre suivant on explore les différents standards définissant les types de codage du signal parole.

Chapitre 3

Codage de la parole et les effets sur le système de RSR

Sommaire

III.1 Introduction

L'accès à un serveur vocal est fait à travers le réseau téléphonique classique, par les réseaux sans fil ou les réseaux IP (Internet Protocole). Les principaux facteurs qui déterminent la qualité de la voix sont le choix du codec (codeur/décodeur), la perte de paquets. Le nombre de codecs standard développé pour compresser les données de la parole a été rapidement augmenté.

Ce chapitre est en relation avec les problèmes de codage de la parole et pertes de paquets qui se produit dans la reconnaissance du locuteur à travers le réseau. La parole codée transmise à partir d'un terminal client vers un serveur de reconnaissance. Au cours de ce chapitre, on décrit certaines normes de codage de la parole couramment utilisée à savoir les codeurs de la forme d'onde, Codeur paramétrique et codeur hybrides.

On termine ce chapitre par une proposition d'une architecture d'un system de reconnaissance du locuteur à distance (RSR) et étudiant les effets des codecs sur ce system, en tenant compte de trois types de codec de la parole : PCM, DPCM et ADPCM conformément à la norme ITU-T utilisé en téléphonie et VoIP (Voix over Internet Protocole). En outre, pour améliorer les performances de la reconnaissance du locuteur dans un environnement bruyant, nous proposons un nouvel algorithme de détection d'activité vocale (Speech Activity Detection-SAD) en utilisant un seuil adaptatif.

III.2 Techniques de codage de signal parole

Le but de codage du signal parole est de représenter le signal par le moins de bits possible, mais maintenir un certain niveau de qualité du signal parole décodé. En général, les techniques de codage sont basées sur la réduction de la redondance du signal, afin que le signal résiduel (après réduction de redondance) puisse être encodé avec moins de bits que l'originale. Afin d'évaluer un codeur de parole, nous devons tenir compte de: débit binaire, qualité de la parole décodée, coût de calcul, retard introduit et robustesse contre l'acoustique et la dégradation de canal.

Les mesures objectives de qualité telles que le SNR peuvent être utiles, mais ils ne peuvent pas refléter la façon dont un signal décodé est perçu. Une mesure courante de la qualité subjective de la parole décodée est l'opinion score moyen (Mean Opinion Score - MOS) (MOS= 1 inacceptable, MOS =5 excellent) [80]. Il est également possible d'appliquer une mesure objective telle que l'estimation perceptive de la qualité de la parole

(Perceptual Estimation of Speech Quality-PESQ, la recommandation ITU-T P.835 [81]), qui prend en compte la façon dont nous percevons la parole. Les codeurs de la parole peuvent être classés en forme d'ondes, paramétriques et codeurs hybrides.

III.2.1 Codeurs de la forme d'onde

Les codeurs de forme d'onde tentent d'obtenir un signal décodé qui reproduit le signal d'entrée d'origine. Ils fonctionnent à des débits moyens et montrent un niveau acceptable de robustesse contre les bruits acoustiques et canal. Ils peuvent être mis en œuvre à la fois dans les domaines temporels et fréquentiels.

III.2.1.1 Codeur de la forme d'onde dans le domaine temporel

Il existe plusieurs types de codeur.

a) PCM (G.711)

Un exemple bien connu est la modulation par Impulsions Codées (MIC ou Pulse Code Modulation- PCM) qui conforment au UIT-T sous la norme G.711 [82]. Codeur utilisé pour le VoIP et la téléphonie numérique. Le VoIP norme décrit deux algorithmes μ-law et A-law. La version μ-law est utilisée principalement en Amérique du Nord. A-law version utilisée dans la plupart des autres pays en dehors de l'Amérique du Nord. Les échantillons de parole sont représentés par des codes binaires. La bande passante est de 200-3400 Hz et la fréquence d'échantillonnage à 8 kHz. Les deux algorithmes utilisant 8 bits par échantillon qui fournit 50% de réduction de l'utilisation de la bande passante pour signal original échantillonné avec 16 bits et de fréquence d'échantillonnage de 8 kHz, soit une réduction de 128 kb/s à 64 kb/s. MOS est plus de 4 [82].

b) DPCM (G.727)

Modulation par Impulsion Codée Différentielle (MICD ou Differential pulse-code modulation-DPCM) a été normalisée par l'UIT-T sous G.727 [83, 84]. Le taux de fonctionnement typique des systèmes utilisant cette technique est supérieur à 2 bits par échantillon résultant en des taux de 16 Kbits/s ou plus. Le DPCM est un codeur de signal qui utilise la ligne de base de modulation par impulsions codées (PCM), mais ajoute des fonctionnalités sur la base de la prédiction des échantillons du signal. L'entrée peut être un signal analogique ou un signal numérique. La figure III.1 représente le codeur DPCM.

Le codeur MICD transmet une version quantifiée $\hat{e}(n)=Q[e(n)]$ de l'erreur de prédiction ou comme résidu [84]:

$$e(n)=s(n)-\bar{s}(n) \tag{III.1}$$

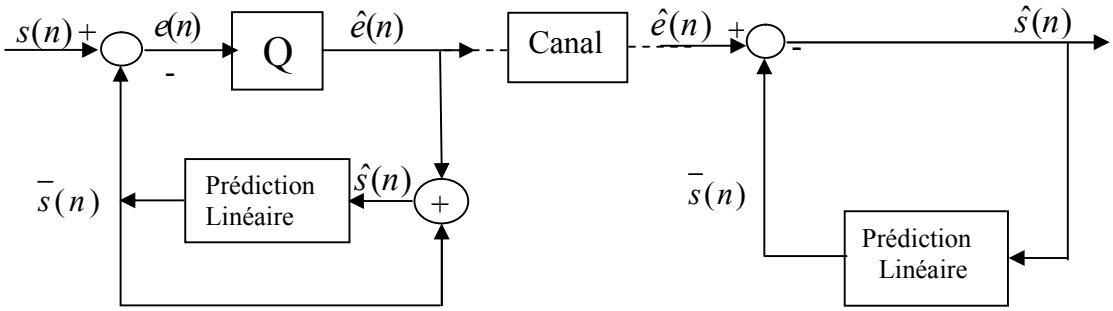


Figure III.1 Codeur et décodeur DPCM.

Où $\bar{s}(n)$ est une prédiction linéaire du signal reconstruit $\hat{s}(n)$ [84]:

$$\bar{s}(n)=\sum_{k=1}^{P} a_k \cdot \hat{s}(n-k) \tag{III.2}$$

Comme déduit de la figure, le signal reconstruit est [84]:

$$\hat{s}(n)=\bar{s}(n)-\hat{e}(n) \tag{III.3}$$

L'opération de l'équation (III.3) s'effectue également au niveau du décodeur, comme le montre la figure III.1. Si le signal d'entrée est suffisamment corrélé, l'erreur de prédiction aura un aspect aléatoire (peu de corrélation) et une gamme dynamique plus petite que le signal d'entrée, de sorte que moins de bits seront nécessaires pour sa quantification. C'est un exemple de la façon d'éliminer les redondances du signal parole pour une transmission plus efficace. Afin d'obtenir des résultats acceptables, un ordre de prédiction non supérieur à 3 est couramment utilisé [15].

c) ADPCM (G.727)

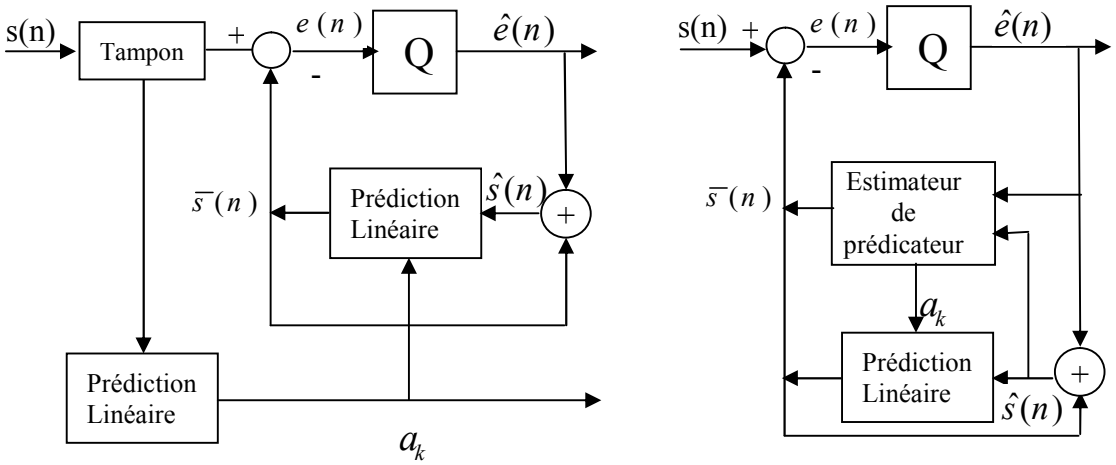
ADPCM (Adaptive Differential pulse Code modulation en français modulation de code impulsionnel différentiel adaptatif) a été normalisée par l'UIT-T sous G.726. Le codec DPCM peut être assoupli en lançant un quantificateur adaptatif ou un prédicteur linéaire adaptatif. Cela donne la modulation de code impulsionnel différentiel adaptatif (ADPCM) [85]. G.726 [85] algorithme permet la conversion de 64 kb/s de signal vers et à partir de

40, 32, 24 ou 16 kb/s en utilisant (ADPCM). L'application de 40 kb/s est de transporter des signaux non vocaux du modem de données. Les débits les plus couramment utilisés pour la compression de la parole sont de 32 kb/s ce qui signifie le double la capacité par rapport à la G.711. Il y a deux façons d'introduire l'adaptation:

• avant adaptation : l'adaptation est réalisée à la fin de l'émetteur. Ce régime exige la transmission des paramètres d'adaptation (coefficients de prédiction linéaire ou de niveau de signal) comme information adjacente. Il s'agit de l'introduction d'un retard.

• adaptation en arrière : les paramètres d'adaptation sont prélevés sur le signal décodé lui-même, où il n'est pas nécessaire d'envoyer les informations du côté.

La figure III.2 montre les codeurs ADPCM, correspondant à une adaptation en avance et en arrière des coefficients de prédiction linéaire.



a) Adaptation en avant

b) Adaptation en arrière

Figure III.2 Avant et en arrière des codeurs ADPCM [85].

III.2.1.2 Codeur de la forme d'onde dans le domaine fréquentiel

Les codeurs dans le domaine fréquentiel sont basés sur la décomposition du signal en composantes de fréquences qui peuvent être quantifiées et codées de façon indépendante. Les principaux types sont les codeurs sous-bande et transforment codeurs [86].

Les codeurs Sous-bande (Figure III.3) utilisent l'analyse de banc de filtre pour décomposer le signal d'entrée x(n) en bandes de fréquences. Le décodeur utilise une banque de filtres de synthèse pour fournir un signal reconstruit y(n) [86]. L'augmentation du nombre de signaux est compensée par la décimation. Par exemple, une banque de filtre

fournissant M bandes avec la même bande passante peut utiliser un facteur de décimation à M comme le montre la figure III.3.

Un codeur de sous-bande fournit reconstruction parfaite lorsque [86]:

$$Y(z) = c \times z^{-k} X(z) \tag{III.4}$$

Reconstruction parfaite peut être obtenue en utilisant des filtres avec des transitions nettes (sharp) dans les limites des bandes. Toutefois, cela pourrait donner des lacunes ou des chevauchements dans ses limites. Aussi, fort de filtres impliquent généralement un coût de calcul plus élevé. Dans le cas M = 2, une solution intelligente est l'utilisation de filtres de miroir en quadrature (Quadrature Mirror Filters (QMF)). Deux filtres h0(n) et h1(n) forment une paire QMF si [86]:

$$h_1(n) = (-1)^{-n} h_0(z) \tag{III.5}$$

et donc : $$H_1(w) = H_0(w + \pi) \tag{III.6}$$

Une paire de filtres QMF atteint reconstruction parfaite si [15]:

1. l'ordre du filtre est encore;

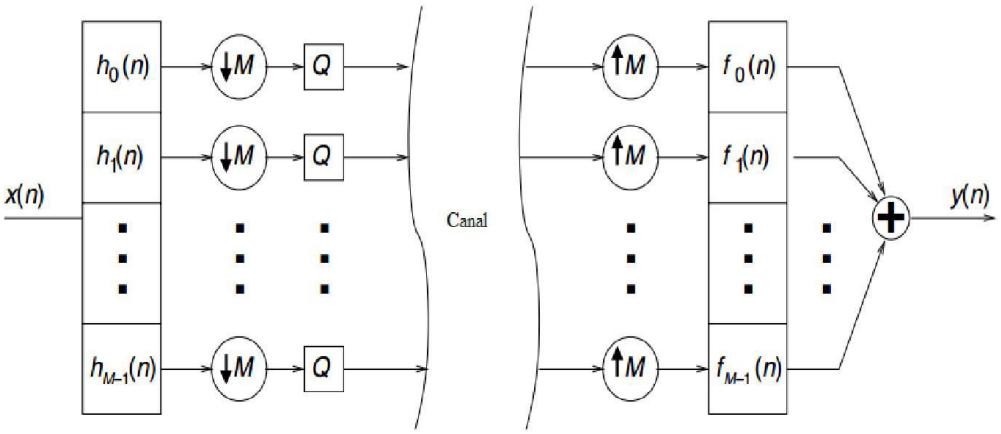


Figure III.3 Schéma général d'un codeur sous-bande [86].

2. ils accomplissent la condition suivante [86]:

$$|H\,0\,(\Omega)|^2 + |H\,1\,(\Omega)|^2 = 1 \tag{III.7}$$

3. le banc de filtres de synthèse doit satisfaire [86]:

$$F_0(z) = H_0(z) \tag{III.8}$$

$$F_1(z) = -H_0(-z) \tag{III.9}$$

Une paire de filtres QMF est des filtres à phase linéaire FIR, utilisée dans le codeur sous-bande UIT-T G.722 (UIT-T, 1988b). La largeur de bande de G.722 est 50-7000 Hz, et le débit est de 64 kbps (56 et 48 kbps sont également disponibles). Il utilise deux quantificateurs ADPCM (faible bande passante de 48 kbps; haute bande passante de 16 kbps). G.722 est approprié pour la vidéoconférence et les applications RNIS [15].

III.2.2 Codeurs paramétriques

Codeurs paramétriques ou vocodeurs, sont basés sur le modèle de production de la parole. Ce modèle est représenté par un ensemble de paramètres actualisés périodiquement. Pour déterminer ces paramètres, le signal est segmenté à intervalles périodiques, appelé *trame*. Les paramètres sont généralement mis à jour à chaque trame. Ils travaillent sur une base trame par trame, de sorte qu'il est nécessaire uniquement pour transmettre les paramètres de modèle qui correspondent à la trame de signal actuel. Au niveau du récepteur, le modèle est utilisé pour synthétiser le signal parole à partir des paramètres reçus. Contrairement aux codeurs de forme d'onde, codeurs paramétriques ne cherchent pas à reproduire le signal d'entrée de forme d'onde, mais pour obtenir un signal de sortie codec qui est la perception similaire à celui d'origine.

Le modèle le plus couramment utilisé est le modèle LPC déjà vu dans le chapitre 1 (I.6.1.2). Ce modèle suppose que le signal de parole soit la sortie d'un filtre numérique qui représente le conduit vocal (Equation I.5), la figure III.4 représente le modèle de conduit vocal d'un signal parole.

La Figure III.4 représente le modèle de la production de la parole généralement adopté pour créer artificiellement des sons comporte :

– un générateur périodique d'impulsions unité ;
– un générateur de nombres aléatoires à valeur moyenne nulle et variance unité
– un commutateur servant à choisir les sons voisés ou non
– un gain proportionnel à la valeur efficace du signal $s[n]$
– un filtre tous pôles $H(z) = 1/A(z)$.

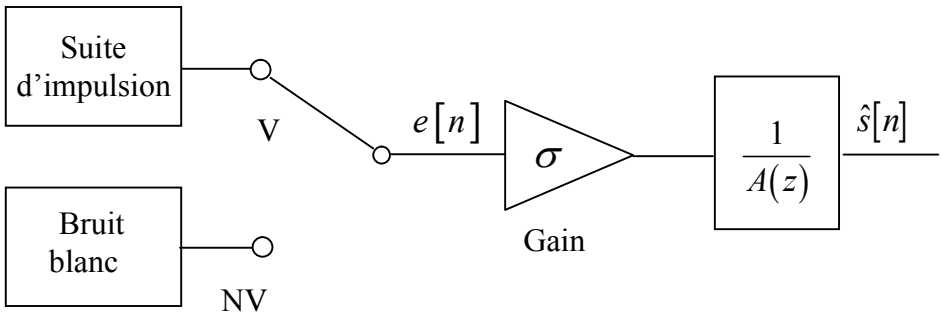


Figure III.4 Modèle du conduit vocal (V=Voisé, NV=Non-voisé).

L'extraction de générateur et du filtre est faite pour chaque trame (la trame est de 20 à 30 ms). Le résidu e[n] de la prédiction linéaire peut être considéré comme le signal d'excitation servant à créer le signal s[n] en passant à travers le filtre récursif (Équation I.5). Ce codage paramétrique (LPC) nécessite la transmission des paramètres de filtre (coefficients LPC et le gain), une décision voisée/non voisée et une estimation de la fréquence fondamentale dans les parties voisées du signal parole à traiter.

Le standard 1015 LPC10e (Tremain, 1982 [87]) utilise un codeur LPC paramétrique avec un taux d'échantillonnage de 8 khz et 22,5 ms longueur de trame. LPC10 [88] utilise un nombre de paramètre p = 10 et 54 bits/tramés, et fournissant un débit final de 2,4 kbps. Avant l'émission, les coefficients LPC sont transformés en rapports log-région (A1 et A2) et coefficients de réflexion (a3 à à0). Dans le cas d'une zone voisée, un bit est utilisé pour la décision voisée/non voisée, six bits pour la fréquence fondamentale, cinq bits pour le gain, cinq bits pour chacun des quatre premiers coefficients du conduit vocal, quatre bits pour des coefficients de 5 à 8, trois bits pour le neuvième et le dixième de deux et un bit de synchronisation [89]. Dans le cas d'un son non voisé, seuls les quatre premiers coefficients sont transmis, et les bits libres sont utilisés pour la protection des erreurs de canal. Ainsi, la LPC10 peut être considérée comme un codeur débit variable. Bien que son intelligibilité soit acceptable, il fournit un score MOS aussi bas que 2.2.

Le codeur de LPC10 a été remplacé par la norme fédérale (Mixed Excitation Linear Prediction MELP [90]) vocodeur à 2,4 kbps, avec un MOS de 3,3 (McCree 1995 [91]).

III.2.3 Fréquence fondamentale (Pitch)

Les vocodeurs LPC précédents et d'autres codeurs de parole que celles étudiées dans la section suivante exigent une estimation de la fréquence de hauteur. Par exemple, la norme LPC10 utilise une fonction de différence d'amplitude moyenne AMDF (Average Magnitude Difference Function). Dans la littérature il ya plusieurs méthodes de détection du pitch que l'on peut diviser en quatre catégories :

- méthodes temporelles, on cite: méthode d'autocorrélation [92], AMDF [92], ASDF (Average Square Difference Function) [93], SIFT [94] (Simplified Inverse Filtering)…….
- méthodes fréquentielles, on cite: Cepsral, HPS [95] (Harmonic Product Spectrum), SHS (Sub-harmonic Summation) [96]……
- méthodes temps-fréquence, on cite: Spectrogramme, Wigner-Ville, pseudo Wigner-Ville lissé, Choï-Williams, Zhao-Atlas-Marks, scalogramme (ondelettes continues) et Wigner-Ville lissé [97].
- temps-échelle, on cite: C'est la détection de pitch par ondelettes.

III.2.3.1 Estimation de la fréquence fondamentale

Pour le cas de la méthode d'autocorrélation, la fréquence fondamentale s'obtient par le maximum de la fonction d'autocorrelation du signal parole. La valeur du fondamental estimée est alors [92]:

$$F_0 = \frac{Fs}{\max} \qquad \text{(III.10)}$$

Où Fs: est la fréquence d'échantillonnage

max: est l'échantillon qui correspond au maximum.

Pour la méthode d'AMDF, le raisonnement pour le calcul de la période du fondamental est analogue au raisonnement pour l'autocorrélation. En effet, a la fonction AMDF appliquée à un signal périodique présent des minima aux multiples de la période du fondamental. Le premier minimum correspond donc à la période du fondamental. La recherche du minimum de cette fonction permet de connaître la période de fondamentale. La valeur du fondamental estimé est alors (Équation III.10). La fonction de la différence de la magnitude définie par [92]:

$$\mathrm{AMDF}(\mathrm{m}) = \frac{1}{\mathrm{L}} \sum_{n=1}^{\mathrm{L}} \left|\mathrm{s}(\mathrm{n}) - \mathrm{s}(\mathrm{n}+\mathrm{m})\right| \qquad 0 \le \mathrm{m} \le n-1 \qquad \text{(III.11)}$$

L : est la longueur de la fenêtre choisie.

m : est le coefficient glissant de la fenêtre.

La méthode Average Square Difference Function (ASDF), est une l'une des méthodes d'estimation de la période à l'aide *des Fonctions de Différences Moyennées*:

$$\mathrm{ASDF[m]}=\frac{1}{N-m}\sum_{n=0}^{N-1-m}(x[n]-x[n+m])^2 \qquad \text{(III.12)}$$

N : nombre d'échantillons.

m : est la longueur de la fenêtre d'analyse.

Pour la méthode de cepstre, le signal périodique peut être considère comme la convolution d'un train d'impulsions par un filtre amorti 'H' Dans le domaine des fréquences, les spectres sont multipliés mais en prenant le « log » du résultat, on obtient la somme des résultats. Une convolution dans l'espace des temps correspond à une addition dans le domaine du cepstre: Si les deux spectres ont des caractéristiques différentes, il devient possible de les séparer. Le cepstre du filtre 'H' a un support temporel localisé au voisinage de 0, alors que celui de la source 'S'est un peigne d'impulsions à la période P et à décroissance lente. On peut donc estimer la période P en déterminant l'intervalle de temps qui sépare deux impulsions successives, ou en recherchant le maximum global du cepstre sur un intervalle n Є $[n_{min}, n_{max}]$, avec nmin > 0 [94].

III.2.4 Codeurs Hybrides

Codeurs hybrides peuvent être considérés comme un mélange de forme d'onde et les codeurs paramétriques. Ils sont paramétriques dans le sens où ils utilisent un modèle paramétrique, mais aussi essayer de conserver la forme d'onde de signal. Les techniques de codage hybrides utilisant des méthodes d'encodage de la forme d'onde et prenant en compte certaines propriétés de la parole ou de la perception auditive. Le principal représentant de cette classe est le codage CELP (Code excited Linear prediction) [98]. Codeurs hybrides sont généralement complexes et ont une plage moyenne de débits de 4 à 16 kbps [61] et ont une meilleure qualité par rapport à celui obtenu à partir de la forme d'onde de codage avec des taux plus élevés [61].

On distingue en général 4 plages de débits pour les codeurs de la parole:

- les hauts débits, supérieurs à 16 kbit/s, correspondant à des algorithmes de codage de la forme d'onde non spécifiques à la parole,
- les débits moyens, de 4 kbit/s à 16 kbit/s, le principal représentant de cette classe est le codage CELP (Code Excited Linear Prediction).
- les bas débits, de 600 bps à 4 kbit/s, correspondant aux codeurs paramétriques appelés aussi vocodeurs (VOICE CODER) spécifiques au codage de la parole.

- les très bas débits inférieurs à 600 bps.

La plupart des nouveaux réseaux de communications et en particulier les réseaux mobiles utilisent des codeurs du type hybride comme les codeurs CELP [61]. Si un modèle LPC est considéré, cet objectif peut être atteint par l'amélioration de l'excitation $\hat{u}(n)$ (ou, de manière équivalente, $\hat{e}(n) = \sigma \hat{u}(n)$, σ représente le gain) utilisé par le modèle de LPC de base [15]. Le point clé est maintenant de savoir comment déterminer les paramètres optimaux de la nouvelle excitation $\hat{e}(n)$, une fois un modèle raisonnable a été choisi pour elle. Ceci est habituellement fait en utilisant la procédure d'analyse par synthèse ABS (analysis-by-synthesis) [15] représentée sur la figure III.5. Elle consiste à obtenir les paramètres qui permettent d'atteindre le meilleur ajustement entre l'original $s(n)$ et des signaux synthétisés $\hat{s}(n)$. Cela peut être fait par l'application d'un critère d'erreur quadratique moyenne minimale (Minimum Mean Square Error-MMSE). L'erreur quadratique moyenne (Mean Square Error) pour être minimisée est calculée comme [99] :

$$E = \sum_{n=0}^{N-1} \left(s(n) - \hat{s}(n) \right)^2 \qquad \text{(III.13)}$$

Le trait le plus caractéristique dans l'architecture de l'ABS est que le codeur contient le décodeur, qui fournit les connaissances nécessaires de $\hat{s}(n)$ du signal décodé. La structure de la boucle du système AbS indique que l'optimisation est effectuée de manière itérative. De la même manière que pour le codeur paramétrique, l'information transmise comprend enfin les paramètres de l'appareil et d'excitation vocale [99].

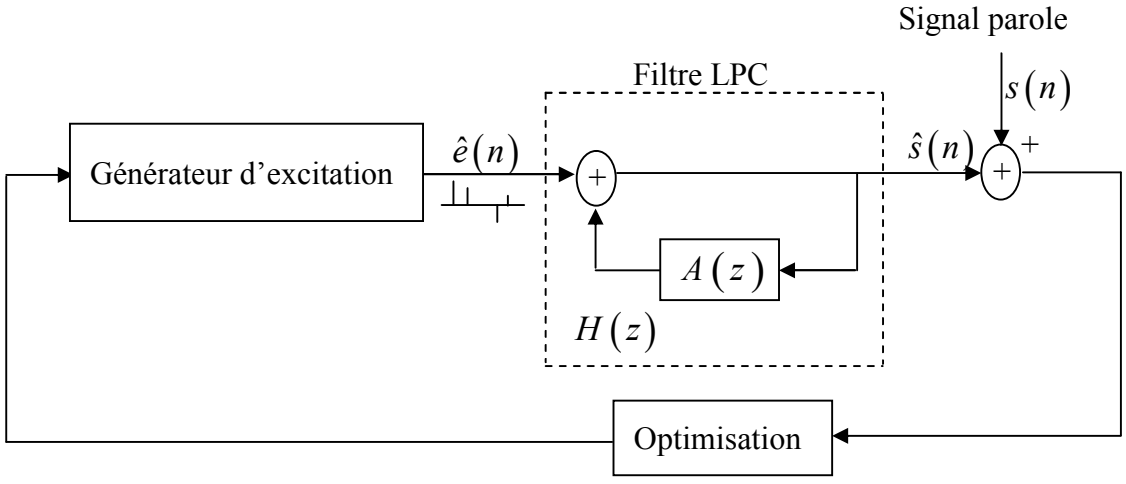


Figure III.5 Procédure d'analyse par synthèse ABS (Analysis-By-Synthesis) [99].

III.2.4.1 Codeur à impulsions multiples

Dans le codage à impulsion multiple (Atal et Remde, [99]), une excitation $\hat{e}(n)$ est construite et consiste en une série d'impulsions l avec des amplitudes b_k et les positions $n_k \ (k = 1, \ldots\ldots l)$ [99] :

$$\hat{e}(n) = \sum_{k=0}^{l-1} b_k \delta(n - n_k) \tag{III.14}$$

Où δ (n) est la fonction d'impulsion unitaire (Impulsion de Dirac). Cette excitation est représentée à la figure III.6. Comme mentionné précédemment, une fois que le modèle d'excitation a été choisi, les paramètres d'excitation (les amplitudes et les positions) doivent être calculés. Ils sont obtenus par minimisation de l'expression [99] :

$$E = \sum_{n=0}^{N-1} \left(s(n) - h(n) * \hat{e}(n) \right)^2 \tag{III.15}$$

Un cas particulier de codage multipulse est le codeur de l'excitation d'impulsion régulière (Regular Pulse Excitation-RPE), qui a été largement utilisé pour la transmission de la parole sur la Transcanadienne à taux plein (Full rate- FR) pour le GSM [61]. Ce codeur est communément connu comme GSM-FR (spécification ETSI GSM 06.10) et utilise une technique RPE avec prédiction à long terme - LTP (long-term prediction) [15].

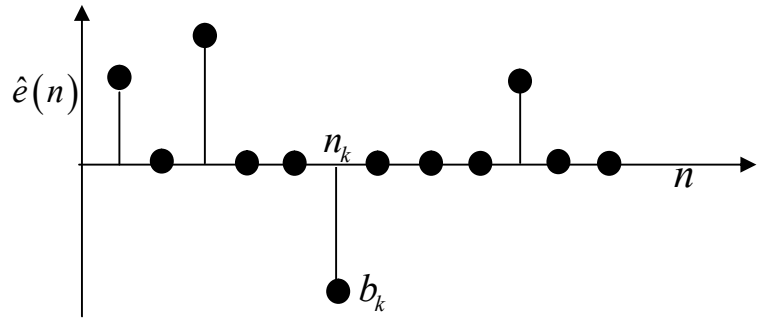


Figure III.6 Excitation de codeur à impulsion multiple [15].

III.2.4.2 Codeur CELP

L'innovation de CELP (Code Excited Linear Prediction) codeurs est de fournir un dictionnaire contenant un ensemble vaste et fixe des excitations $c_k(n)$ (k = 0, M -... 1)

(Schroeder et Atal, 1985 [98]). La formulation de base du codeur CELP suppose que les excitations M sont des séquences aléatoires gaussiennes. L'idée de CELP est représentée dans la figure III.7. L'excitation est construite comme suit [98]:

$$\hat{v}(n) = \gamma_k \cdot c_k(n) \qquad \text{(III.16)}$$

où γ_k est obtenu au cours de la procédure d'optimisation avec $c_k(n)$.

Le schéma CELP ne nécessite que l'indice correspondant à la meilleure entrée de dictionnaire de codes et le gain pour coder le résiduel à long terme $\hat{v}(n)$.

Un exemple de CELP est CELP à faible délai de 16 kbps normalisé par l'UIT-T G.728 (Chen, 1990 [100]), ce codeur atteint 16 kbps avec un débit de seulement 2 ms. Ceci peut être accompli en utilisant un prédicteur à court terme (STP- Short Term Predictor) avec une adaptation en arrière et d'éviter l'utilisation d'un prédicteur à long terme (LTP- long-term predictor), qui est pallié en utilisant une commande de STP de 50. La trame est de 2,5 ms de longueur 20 échantillons et comprend 4 sous-trames de 5 échantillons. Le dictionnaire utilise 7 bits des mots de code mis en forme par un gain de 3 bits.

La technique CELP a également été dérivée dans un certain nombre de variantes telles que: la somme de vecteur de prédiction linéaire excités VSELP (Vector Sum Excited Linear Prediction) adoptée par le IS-54/136 (7,95 kbps) et GSM (pour la moitié du taux (Half Rate -HR)) canal, 5,6 kbps), les systèmes mobiles (Qualcomm Code Excited Linear Prediction) QCELP adoptée par IS-95 (8.5-4-2-0.8 kbps, variable selon l'activité vocale) et le (Algébrique CELP, ACELP) [66]. Cette dernière variante (ACELP) a été largement appliquée au cours des dernières années dans les réseaux mobiles et IP. Le paragraphe suivant est consacré [66].

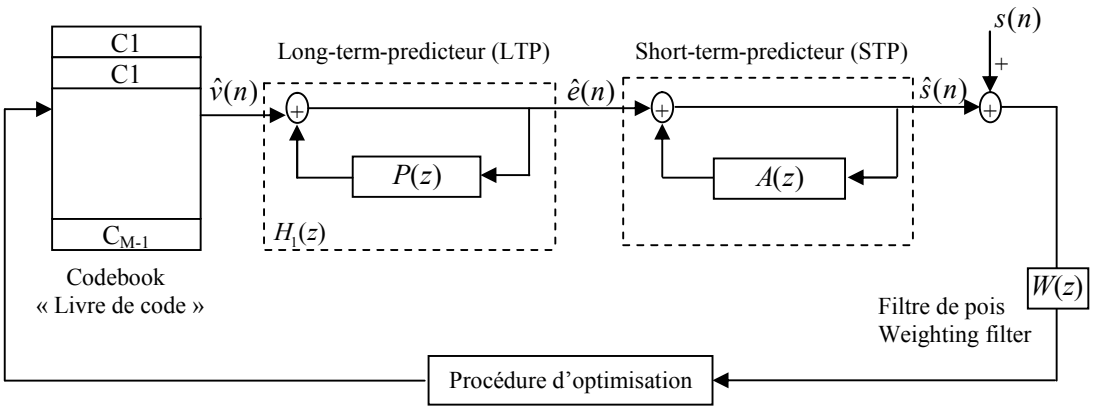


Figure III.7 Schéma général de codeur CELP [66].

III.2.4.3 Codeurs ACELP

Le CELP algébrique (ACELP) est une modification du codeur CELP, dans lequel les mots de code sont des codes algébriques creux inventés par (Laflamme 1990 [101]). Ces codes contiennent principalement des zéros, ce qui permet une recherche de livre de codes rapide. Codes Dispersés sont obtenus à partir de codes de permutation, qui sont obtenus en permutant une série d'impulsions dans une série de positions préfixées [101].

Un exemple important d'un codeur ACELP est le codeur GSM EFR [61]. Un schéma de principe du décodeur est représenté sur la figure III.8. Le débit est de 12,2 kbps avec un MOS d'environ 4.

Les codes algébriques de l'EFR (plein débit amélioré-Enhanced Full rate) sont construits à partir de 5 codes de permutation, qui permettent le positionnement des 10 impulsions (2 impulsions par code de valeurs ±1) dans chaque sou-trame. Les positions des impulsions sont déterminées selon un critère d'AbS pour réduire au minimum l'erreur perceptuellement pondérée.

Une autre norme ressemble beaucoup à l'EFR est le codeur de IS-641 employé par le système cellulaire de IS-136. Son débit est de 7,4 kbps complété avec 5,6 Kbits/s du de codage de canal, ce qui donne un total de 13 kbit/s (Honkanen et coll., 1997). Le système de téléphonie de IS-95 admet également l'utilisation d'un codeur RCELP (CELP détendue) avec une recherche de livre de codes algébriques. Ce codeur est connu comme codeur amélioré à taux variable (EVRC-Enhanced Variable Rate Coder (EVRC),) et permet un fonctionnement de débit variable (8,55, 4 et 0,8 kbit/s) (TR45, 1996).

Les spécifications FR et RFE comprennent la substitution et l'inhibition des mécanismes d'atténuation erreur de canal. Une approche alternative pour éviter l'effet gênant de canal Erreurs consiste à augmenter le nombre de bits consacrés au codage de canal et celles consacrées à source codage afin de maintenir la constante de vitesse d'échantillonnage total. Le GSM – AMR (Adaptive Multirate-AMR) codeur (ETSI, 1998 b) implémente cette idée. En fait, AMR est une famille de codeurs ACELP avec la même structure que l'EFR, mais travaillant à différents débits. À un instant donné, la sélection d'un codeur spécifique dépend de la condition de canal. La vitesse de transmission plus élevée est choisie pour un canal en bon état, alors que celui le plus bas est sélectionné lorsque le canal est gravement détérioré. Les bits restants sont consacrés au codage de canal jusqu'à l'achèvement de la vitesse de transmission total (22,8 kbps pour le canal TCH/F et 11,4 kbit/s pour le canal TCH/H). Les débits AMR sont 12,2 (c'est le codeur EFR), 10.2, 7,95, 7.4, 6,7, 5.9, 5.15 et 4,75 pour le canal TCH/F et 7.40, 6,7, 5.9, 5.15 et

4,75 kbit/s pour le canal TCH/H. Les codeurs AMR utilisent une longueur de trame de 20 ms et un jeu de coefficient LSP est calculé une fois par image (sauf à 12,2 kbit/s, EFR) en appliquant une fenêtre asymétrique qui met l'accent sur les échantillons plus récents. La fiche d'AMR a également été adoptée par l'UMTS (spécification 3GPP TS 26.090).

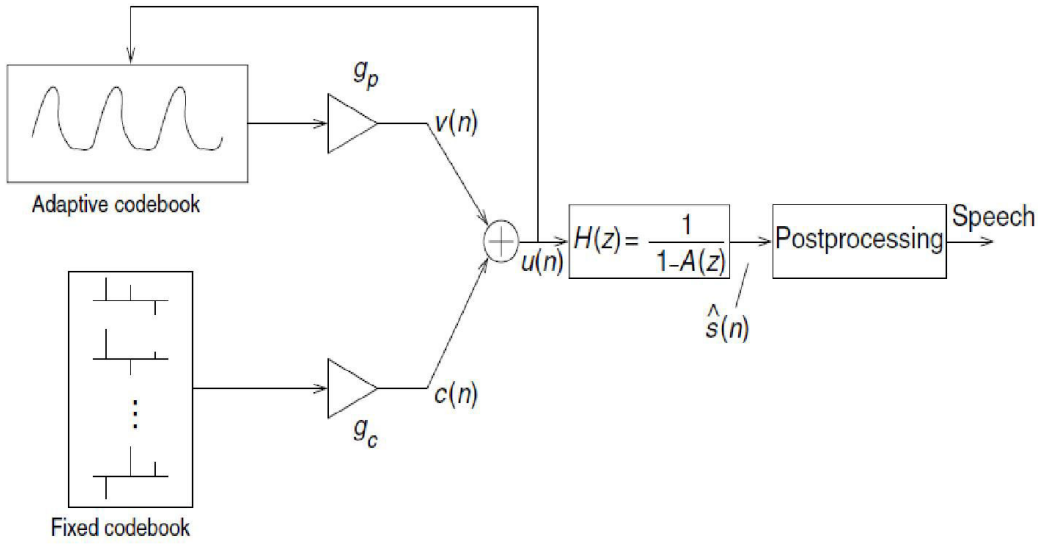


Figure III.8 Diagramme de décodeur GSM-EFR [15].

III.3 Effets de codecs sur un système de RSR en utilisant un nouveau SAD

Dans cette partie on présente une architecture d'un system de reconnaissance à distance (RSR), ainsi on traite les effets des codecs de la parole sur les performances de RSR du mode indépendant du texte dans les applications de VoIP, en tenant compte de trois types de codec: PCM, DPCM et ADPCM conforme à ITU-T (Union Internationale des télécommunications-Telecoms) utilisés en téléphonie et VoIP (Voice over Internet Protocol).

Afin d'améliorer les performances de RSR en environnement bruyant, nous proposons un robuste algorithme de détection des zones de la parole (SAD) à l'aide d'un " seuil adaptatif ", qui peuvent être simulé avec des fichiers parole immergé dans un bruit additif de la base de données TIMIT (Texas Instruments Massachusetts Institute of Technology) [102] qui permet au système de reconnaissance se faire sous des conditions presque idéales. En outre, notre système RSR est basé sur la quantification vectorielle (VQ) et les paramètres MFCC. L'extraction de caractéristiques procédés après (pour la phase de test) et avant (pour la phase de formation) l'envoi de signal parole sur le canal de communication. Par conséquent, les canaux numériques peuvent introduire plusieurs types

de dégradation. Pour remédier à la dégradation du canal, un code convolutif est utilisé comme contrôle d'erreur de codage avec le canal à bruit blanc additif gaussien (AWGN).

Finement, les résultats de simulation donnent des meilleures performances en termes du taux de reconnaissance et temps d'exécution pour le codage par PCM en comparaison avec DPM et ADPCM.

III.3.1 Introduction

Notre objectif est de fournir une évaluation des codecs vocaux, compte tenu de codecs conformes à l'UIT-T et qui sont utilisés dans la téléphonie sur Internet et le réseau IP dans un système de reconnaissance du locuteur distant. Les codecs respectent généralement les normes de l'UIT-T. Les principales normes de l'UIT sont: G.711, G.722, G.723, G.726, G.728, G727 et G729. Nous considérons dans cette section les codecs PCM, ADPCM et DPCM et de leurs effets sur la précision de la reconnaissance du locuteur. Bien qu'il existe deux architectures pour la mise en œuvre d'un système de reconnaissance du locuteur à distance sur un canal numérique. Dans la première approche, généralement connu comme la reconnaissance du locuteur/parole dans le réseau NSR (paragraphe II.2). La Figure II.1 montre un schéma de cette architecture du système. La deuxième approche dite reconnaissance du locuteur/speech distribué DSR (paragraphe II.2) Le schéma conceptuel de la DSR est montré dans la figure II.2.

Dans notre travail on a adopté la RAL conformément à la figure II.1 (NSR).

III.3.2 Configuration du système proposé

Dans la reconnaissance du locuteur on distingue deux tâches principales : l'identification et la vérification. Le système que nous allons décrire est classé en tant que système d'identification du locuteur en mode indépendant du texte à distance et qui a été mis en place selon le diagramme représenté par la figure III.9.

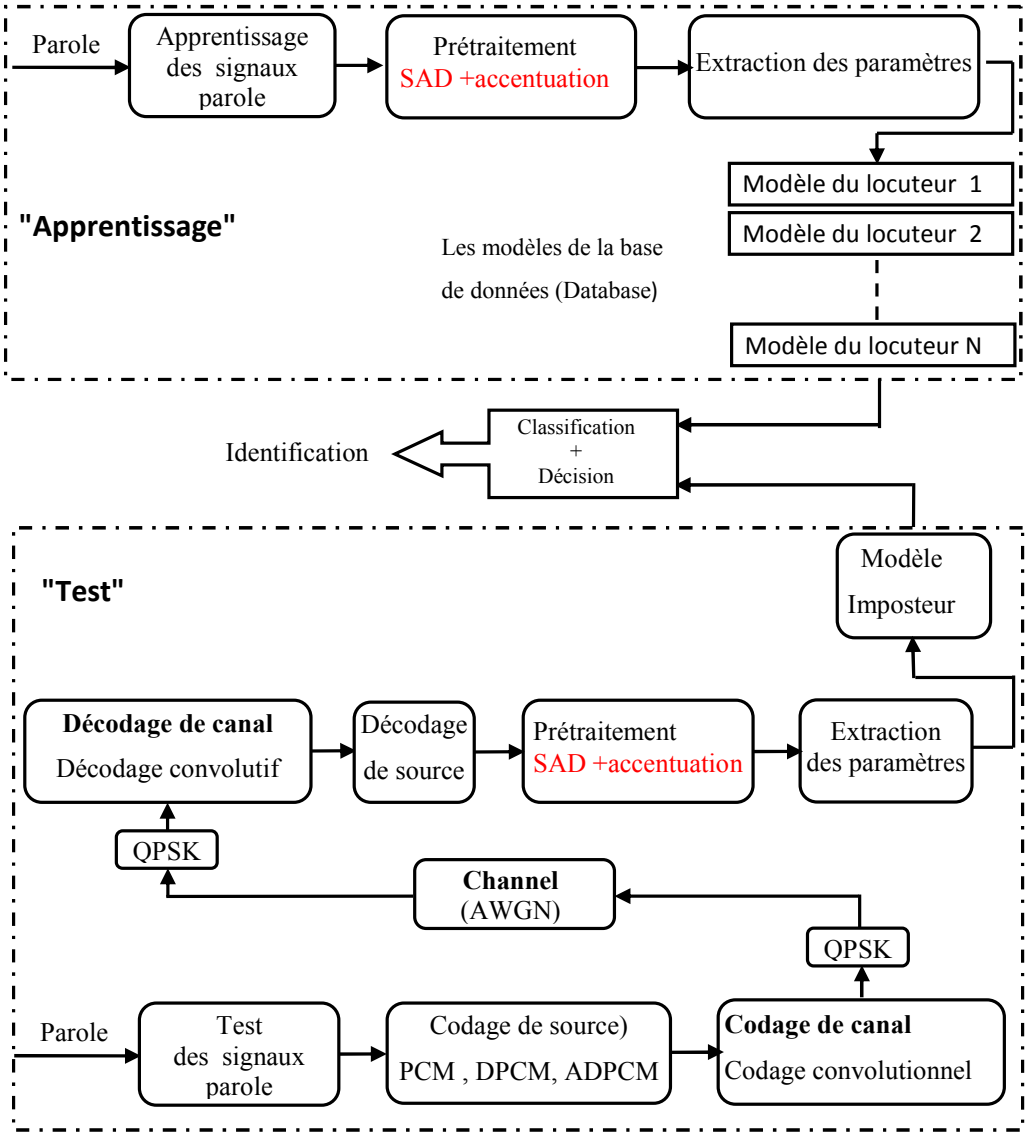


Figure III.9 Schéma du système RSR proposé.

III.3.2.1 Phase d'apprentissage

Dans la phase d'apprentissage, c'est le processus de génération de modèles spécifiques à chaque locuteur avec les données recueillies. Le modèle génératif utilisé dans la reconnaissance de l'orateur est la quantification vectorielle VQ (paragraphe I.6.2).

Le système a été formé en utilisant des signaux parole de la base de données TIMIT où nous avons utilisé 30 locuteurs provenant de différentes régions (10 hommes et 20 femmes). Le signal de parole passait par la phase de prétraitement (accentuation + détection des zones de la parole SAD). Où, nous passons le signal parole par un filtre numérique avec 6 dB / Octave avec un facteur de 0,97 habituellement.

Après avoir accentué le signal parole, des segments de silence sont éliminés par l'algorithme de détection d'activité vocale (SAD) et vingt-quatre coefficients MFCC sont

extraits et forment la caractérisation des modèles en utilisant la quantification vectorielle (VQ).

III.3.2.2 Phase de Test

Dans cette étape, nous avons utilisé des codecs PCM, ADPCM et DPCM et donc leurs coefficients sont convertis en une séquence binaire. Avant d'utiliser la modulation QPSK nous introduisons un code correcteur d'erreurs (FEC) pour mettre le system plus robuste aux erreurs de canal de transmission, le signal codé est transmis à travers le canal AWGN. Après démodulation (QPSK), le décodage de convolutif, et décodage der source (PCM, DPCM, ou ADPCM), les données binaires sont reconverties en un signal parole synthétisée. Enfin, les coefficients MFCC sont extraits (à partir du signal synthétisé).

a) Choix de code correcteur

Il existe différents types de techniques FEC, à savoir le code de Reed-Solomon et des codes de convolutif. L'algorithme de Viterbi est un procédé de décodage des codes convolutifs [103]. Les codes convolutifs, qui sont fréquemment utilisés comme codes de correction d'erreur dans les systèmes de transmission numérique, sont généralement décodés à l'aide du décodeur de Viterbi [104]. Le codage convolutif est effectué en combinant le nombre fixe de bits d'entrée [103]. Les bits d'entrée sont stockés dans le registre à décalage de longueur fixe et ils sont combinés avec l'aide de mod-2 additionneurs. Une séquence d'entrée et le contenu des registres à décalage exercent une addition modulo-deux après séquence d'information est envoyé aux registres à décalage, de sorte qu'une séquence de sortie est obtenue. Cette opération est équivalente à la convolution binaire et d'où il est appelé codage convolutif [103]. Le rapport « R = k / n » est appelé le taux de code d'un code convolutif où « k » est le nombre de bits d'entrées parallèles, et « n » est le nombre de bits décodés parallèle de sortie, le nombre m est signifie les registres à décalage. Registres à décalage stockent les informations d'État du codeur convolutif et longueur de contrainte (K) rapporte au nombre de bits dont dépend la sortie. Un code convolutif peut devenir très compliqué avec différents taux de codage et les longueurs de contrainte [103]. Un simple code de convolutif avec un rapport de code de ½ est représenté par la figure III.10. « m » représente le bit de message actuel, *u1* et *u2* représentent les précédents deux messages successifs de bits stockés et qui représentent l'état du Registre à décalage [103]. Ce taux est (k / n) = 1/2, avec une longueur de

contrainte K = 3. Ici, k est le nombre de bits d'information d'entrée et n'est pas le nombre de bits de sortie parallèle codés à un intervalle de temps [103].

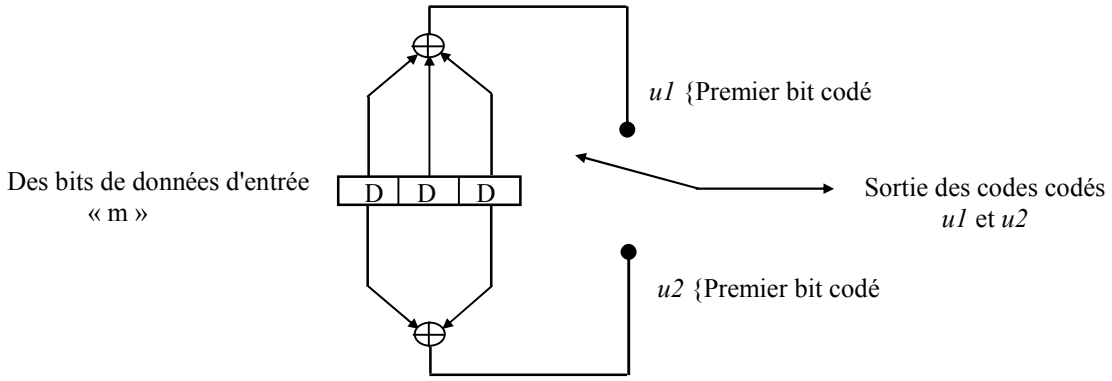


Figure III.10 Codeur convolutif avec ½, où "D" représente le retard (Delay)

III.3.2.3 Phase de Décision

Correspondance de motifs est la tâche de calculer les scores correspondants entre les vecteurs de caractéristique d'entrée (les paramètres d'arrivés de la phase de test) et les modèles donnés [105]. Dans notre travail, nous avons utilisé la méthode de classificateur de Distance euclidienne (DE) [105]. La distance euclidienne classificatrice (DE) a l'avantage de la simplicité et la rapidité de calcul. Le classement se fait par calcul de la distance minimale à fin de décider lequel des locuteurs sur tout l'ensemble d'apprentissage et les plus susceptibles d'être le locuteur de test [105]. Considérez une classe « i » avec « m » composantes moyenne des vecteurs des caractéristiques $\overline{X_i}$, et un vecteur des échantillons Y, donnés respectivement par [105] :

$$\overline{X}_i = \left[\overline{X_{i1}}, \overline{X_{i2}} .... \overline{X_{im}}\right]^t \tag{III.17}$$

Et [105]: $$Y = \left[y_1, y_2 .... y_m\right]^t \tag{III.18}$$

La DE entre la classe i et le vecteur Y est donnée par [105] :

$$d(i,Y) = \left\| \overline{X_{i1}} - Y \right\| = \sum_{k=1}^{m} (\overline{x_{ik}} - yk)^2 \tag{III.19}$$

Pour un certain nombre de classes C, la règle de décision pour le classificateur de ED est que Y être affectées à la classe j si [105]:

$$d(j,Y) = \min \left\{ d(i,Y) \right\}, \forall\, i \in C \tag{III.20}$$

III.3.3 Extraction des paramètres

Les paramètres du locuteur peuvent être extraits en utilisant des coefficients cepstraux (MFCC). Après la préaccentuation et la détection parole /silence, des segments parole sont fenêtrés en utilisant une fenêtre de Hamming. Les calculs des coefficients cepstraux (MFCC) sont montrés dans la figure I.8.

Les signaux paroles de la base de données TIMIT sont échantillonnés avec une fréquence d'échantillonnage Fs de 16000 échantillons/s, ces fichiers ont été sous échantillonnées à une fréquence d'échantillonnage de 8000 échantillons/s. Dans nos expériences de reconnaissance du locuteur, le signal parole est segmenté en trames de 16 ms (128 échantillons), avec un chevauchement (overlapping) de 8 ms (64 échantillons), la transformée de Fourier discrète est pris de ces trames fenêtrées. La grandeur de la transformation de Fourier passe ensuite à travers un filtre-Banque comprenant vingt quatre filtres triangulaires (correspond aux coefficients MFCC).

III.3.4 Algorithme de la détection parole/non-parole (SAD)

La détection de présence de la parole dans un fond de bruit est l'étape importante de prétraitement dans les systèmes d'RAL.

L'élimination des trames qui expriment le silence dans le flux d'entrée (les trames de signal parole) du système de RAL permet de réduire efficacement le taux d'erreur du système. La plupart des SAD effectuent la classification de la parole sur la base des caractéristiques extraites de la trame en cours. Dans notre cas, la détection s'effectue en tenant compte de tous les trames qui représentent le signal parole. Dans la littérature, ils existent de nombreux algorithmes de SAD disponibles [106, 107, 108,109].

Le nouveau SAD est fondé sur deux travaux originaux [110, 111]. Dans [110] l'auteur avait utilisé l'énergie résiduelle LPC et le taux de passage par zéro pour la détection parole/non-parole en utilisant un seuil adaptatif, ce seuil est calculé pour chaque trame introduite en comparaison avec des caractéristiques des trames précédentes, mais l'algorithme suppose que les 15 premières trames soient non-parole, ce qui signifie de probables erreurs. Le deuxième auteur [111], utilise les rapports d'énergies par les passages par zéro pour la classification voisée /non voisée basant sur un seuil fixe.

Notre nouvel algorithme de détection parole/non-parole (SAD) est basé sur les rapports d'énergie par passage par zéro (en anglais : EZR) en utilisant un seuil adaptatif pour détecter les zones d'activité de la parole et de supprimer les intervalles de silence. Le

paramètre présenté EZR [m] est appliqué comme critère de décision parole / non-parole. L'EZR [m] sera par conséquent relativement bas dans les régions de silence du signal de la parole et inversement pour les régions qui expriment la parole.

Le but est alors de préciser tout d'abord avec efficacité et avec minimum de temps si la trame représente un signal parole ou non, si le rapport d'énergie par passage par zéro (EZR) calculé pour une trame supérieure à un seuil, cette trame est alors considérée parole, sinon la trame est considérée comme zone de silence (le système de reconnaissance n'extrait pas de paramètres dans cette région de silence). La figure III.11 exprime le spectre d'un signal émergé dans un bruit additif (bruit de fond), où l'énergie de signal propre est supérieure à celle de bruit de fond.

Notre SAD fonctionne avec une fenêtre glissante d'une forme rectangulaire de 8 ms. La procédure de calcul du seuil est comme suit:

1 - la segmentation du signal de la parole en entier (le signal du locuteur) en trames de 8 ms avec fenêtre rectangulaire et sans chevauchement.

2 - calcul de l'énergie (E [m]) et le taux de passage par zéro (PPZ [m]) pour chaque trame et calculant E [m] / ZCR [m].

3 - calculer le maximum et le minimum d'EZR.

4 - calculer le seuil de decision parole/non-parole (l'équation III).

Le seuil de decision parole/non-parole depend des rapports EZR et de niveau de bruit ("α "), ce seuil estimé par l'algorithme d'une façon automatique et adaptatif. Le rapport EZR est défini par :

$$EZR\left[m\right] = \frac{\overline{E}\left[m\right]}{ZCR\left[m\right]} \tag{III.21}$$

Où ZCR [m] et $\overline{E}[m]$ présentent respectivement le taux de passage par zéro et l'énergie moyenne d'une trame (équation I.1). Le passage par zero (ZCR) est défini par [110]:

$$ZCR\left(m\right) = \sum_{n=0}^{N-1} \left|\mathrm{sgn}\left[x\left(n\right)\right] - \mathrm{sgn}\left[x\left(n-1\right)\right]\right| w\left(m-n\right) \tag{III.22}$$

Où: sgn (.) est la function signe définie par [110]:

$$\mathrm{sgn}\left[x\left(n\right)\right] = \begin{cases} +1, & x\left(n\right) \geq 0 \\ -1, & x\left(n\right) < 0 \end{cases} \tag{III.23}$$

Notre algorithme (SAD) calcule les EZRs de tous les trames (pour le signal locuteur) et estime le seuil (threshold) de detection parole/non-parole par:

$$Seuil = \min\left(EZR\right) + \alpha \times \left[DELTA\right] \qquad \text{(III.24)}$$

Après segmentation de signal parole en différentes trames, on calcul “DELTA” qui représente la difference entre le maximum d’EZR (maxEZR) et le minimum d’EZR (minEZR), DELTA se calcule par:

$$DELTA = \max\left(EZR\right) - \min\left(EZR\right) \qquad \text{(III.25)}$$

α: est un nombre réel dans l'intervalle de] 0,1 [. Dans notre simulation, nous avons fixé α = 0,35 pour different niveau de bruit SNR. La procedure de calcul de seuil de décision praole/non-parole (threshold) est representé par la figure III.12.

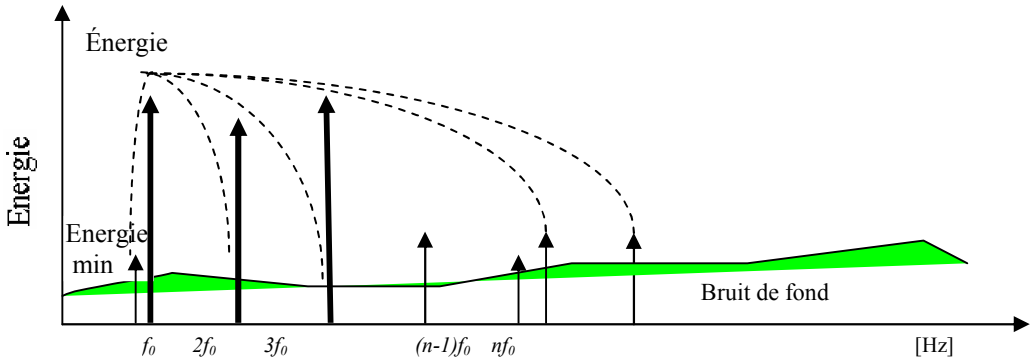


Figure III.11 Illustration d’énergie pour un signal parole, f0 exprime la fréquence fondamentale du signal, 2f0….nf0 exprimes les formants.

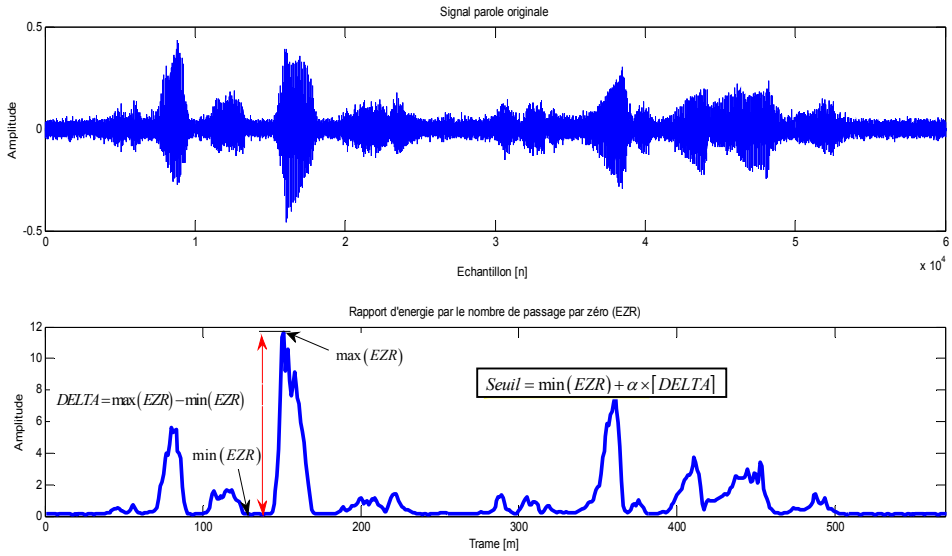


Figure III.12 Procédure de calcul de seuil de décision parole/non-parole, par éstimations d’EZR pour chaque trame.

Nous résumons notre algorithme de détection d'activité de la parole par la figure III.13.

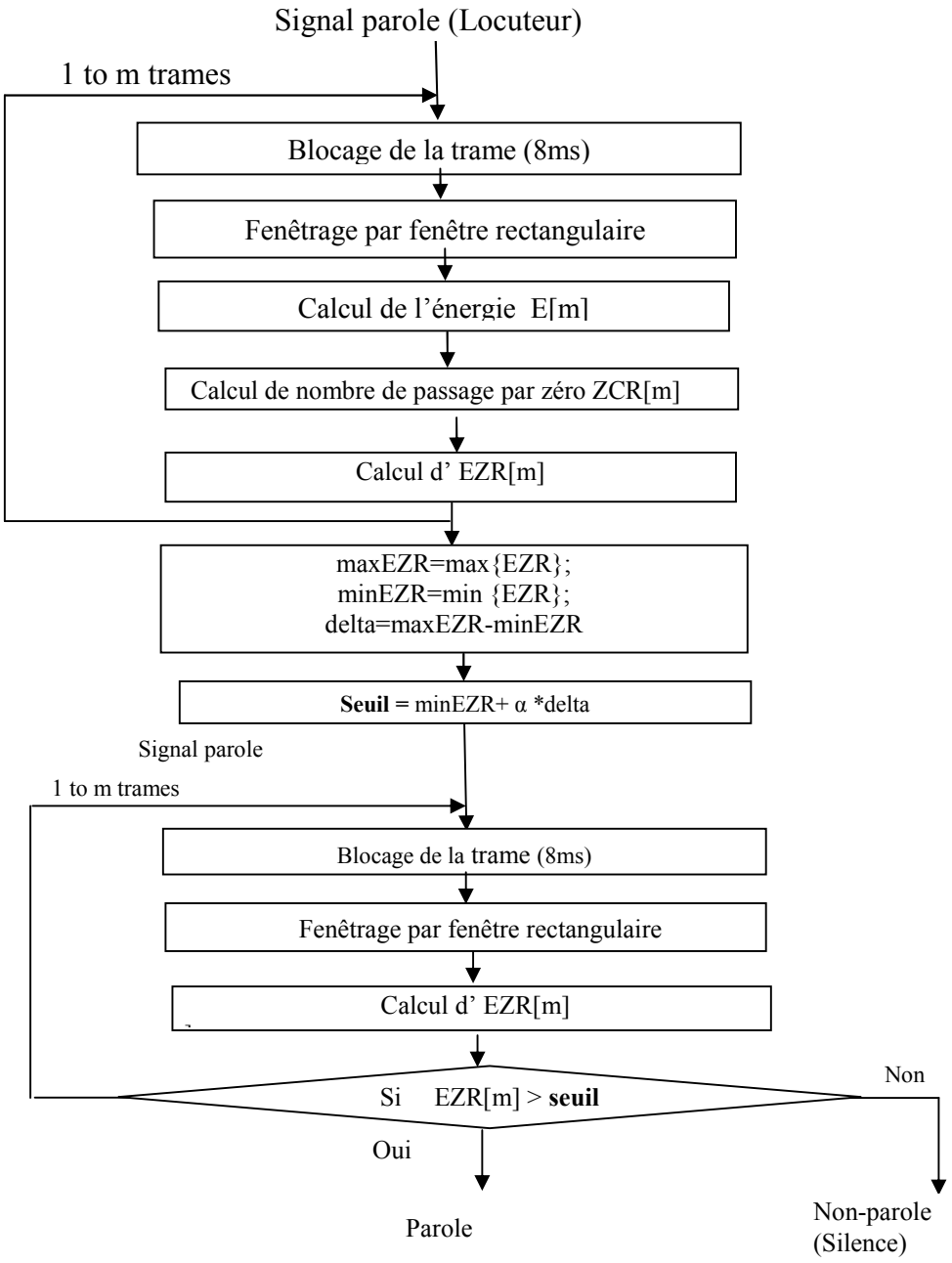


Figure III.13 Algorithme général de detection parole/non-parole.

III.3.5 Résultats de simulation et discussion

Les expériences de l'identification du locuteur sont menées sur la base de données TIMIT. Le corpus de TIMIT a été conçu pour fournir des données de parole pour l'acquisition de connaissances acoustique phonétique et pour le développement et l'évaluation de systèmes de reconnaissance automatique de la parole [102]. Bien que la base TIMIT ait été principalement conçue pour la reconnaissance e la parole, elle est largement utilisée dans la RAL, car elle est l'une des rares bases de données avec un nombre relativement important de locuteurs. Elle contient 630 messages vocaux des

locuteurs (438 mâles et 192 femelles), et chaque intervenant (Locuteur) lit 10 phrases différentes. Dans nos expériences, nous avons choisi 30 locuteurs (10 hommes et 20 femmes). En outre, dans l'étape d'apprentissage, on a utilisé trois énoncés pour chaque locuteur. Le modèle génératif utilisé dans le système de reconnaissance est la quantification vectorielle. Nous calculons le vecteur d'extraction des paramètres à partir de 24 coefficients MFCC.

Le premier test, nous évaluons notre algorithme SAD, où le signal de parole qui est " *She had your dark suit in greasy wash water all year*" passait à travers l'algorithme (SAD) ($\alpha = 0, 35$). La figure III.14 représente le signal parole original. La figure III.15 représente le signal parole après avoir été passé à travers l'algorithme SAD. La figure III.16 illustre un signal parole (sans bruit) et son contour d'activité vocale (SAD).

D'après les figures III.15 et III.16, nous observons l'éfficacité de notre algorithme où des segments de silence sont éliminés, ce qui signifie que la capacité mémoire du système et le taux de reconnaissance seront améliorées.

Les figures III.17, III.18 et III.19, représentent les contours d'activité vocale en fonction d' SNR pour 10 dB, 5 dB et 0 dB respectivement, les zones de silence sont éliminées avec précision. En outre, la détection d'activité vocale est robuste, dont notre SAD a fonctionné avec précision dans des environnements de faible SNR (au-delà d'SNR de 5 dB).

Pour observer l'effet de notre algorithme d'activité vocale sur le taux d'identification du locuteur, nous avons utilisé notre système de reconnaissance du locuteur avec et sans SAD (non à travers le canal numérique). La figure III.20 montre un taux d'identification avec et sans algorithme de détection d'activité vocale en fonction d'SNR. Il est clairement démontré que cette figure représente une amélioration des taux d'identification lors de l'utilisation de notre algorithme SAD dans des environnements d'SNR basse.

Le deuxième test est sur l'effet des erreurs de canal sur un système de reconnaissance du locuteur à distance (RSR). Nous utilisons donc des fichiers audio originaux et reconstruits après la transmission à travers le canal AWGN en utlisant notre schéma du système RSR proposé. Le tableau III.1 montre les résultats de simulation du taux d'identification à l'aide des signaux paroles originaux et reconstituait (synthétisé après transmission à travers le canal AWGN) dont, nous observons la dégradation du taux d'identification lors de l'utilisation des fichiers reconstruits.

Le troisième test consiste à évaluer le taux d'identification du locuteur à l'aide de: PCM, DPM et ADPCM avec notre système de reconnaissance du locuteur à distance. La figure

III.21 illustre l'effet des codecs PCM, DPCM et ADPCM sur notre système de reconnaissance à travers le canal AWGN. À partir de cette figure on peut conclure que le codec PCM donne de meilleurs résultats en présence de bruits.

Le quatrième critère est le temps d'exécution (run time) de chaque codec utilisé dans notre travail. Le tableau III.2 montre les résultats de simulation des techniques PCM, DPCM et ADPCM en terme du temps d'exécution en utilisant notre système RSR, dont nous pouvons observer que DPCM nécessite plus de temps pour exécuter que PCM et ADPCM, mais la technique PCM nécessite peu temps d'exécution (nous avons utilisé un ordinateur portable Intel (R) Core (TM) i5-3210M CPU @ 2, 5 GHz 2.50GHZ).

Le cinquième test est pour juger, quelle est la technique de codage soit la meilleure? une étude a été faite pour certaines techniques de codage (convolutif, Reed Salomon, et Hamming) sur le taux d'erreur binaire (BER). Les résultats sont présentés en matière de BER vis-à-vis du SNR (rapport signal-bruit). Figure III.22 donne une bonne idée sur laquelle des techniques de codage soit la meilleure, nous pouvons observer que le code convolutif soit la meilleure technique de codage de canal. Bien qu'il soit clair que l'utilisation d'un codage de canal donne de bons résultats,

Le sixième test est concernant l'effet de codage de canal sur le taux d'identification du locuteur. Par conséquent, nous avons évalué le taux d'identification avec et sans code de convolutif (AWGN + Code). La figure III.23 montre le résultat de simulation pour le taux d'identification en fonction d'SNR avec et sans code de convolution. De cette figure on peut conclure que le codage de canal (codage convolutif dans notre cas) est nécessaire pour améliorer le taux d'identification.

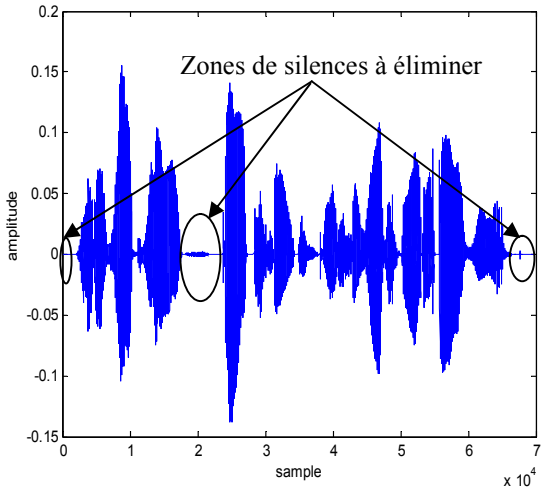


Figure III.14 Signal original

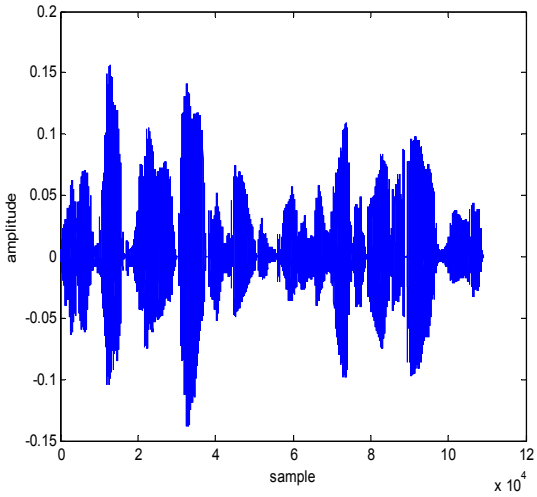


Figure III.15 Signal parole après avoir été passé à travers l'algorithme SAD (α =0.35).

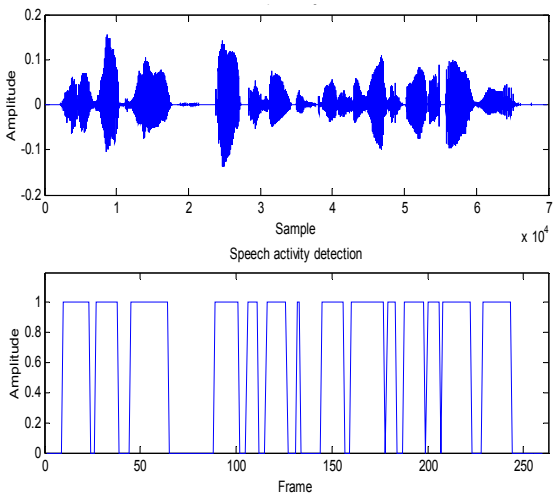


Figure III.16 Signal parole (sans bruit) et son contour d'activité vocale (en bas).

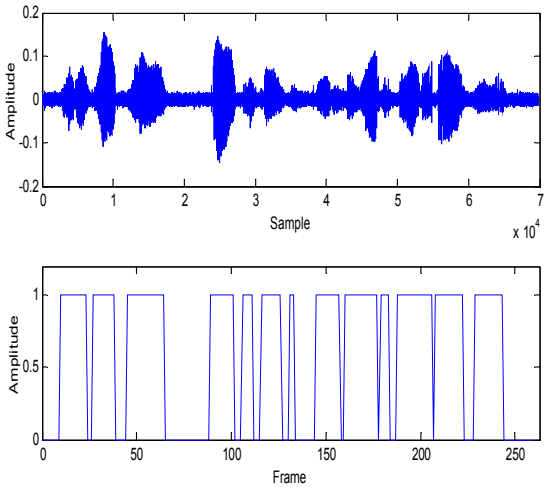


Figure III.17 Signal parole (sans bruit) et son contour d'activité vocale (en bas) à 15 dB.

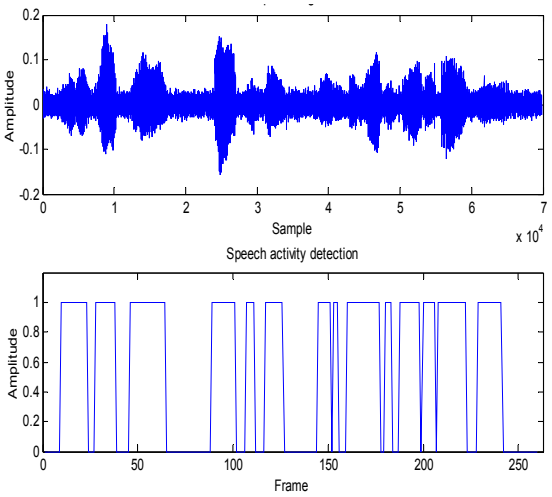


Figure III.18 Signal parole (sans bruit) et son contour d'activité vocale (en bas) à 5dB.

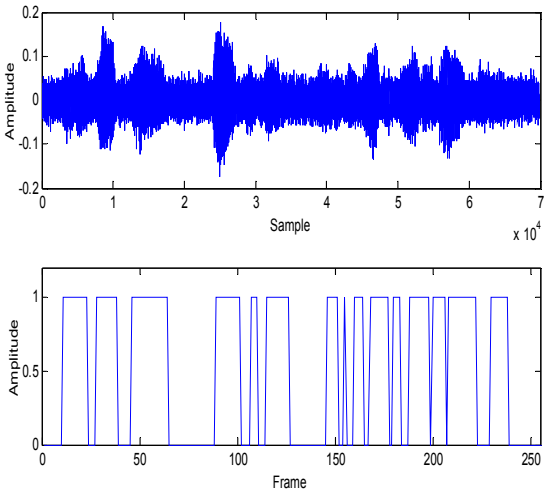


Figure III.19 Signal parole (sans bruit) et son contour d'activité vocale (en bas) à SNR=0dB.

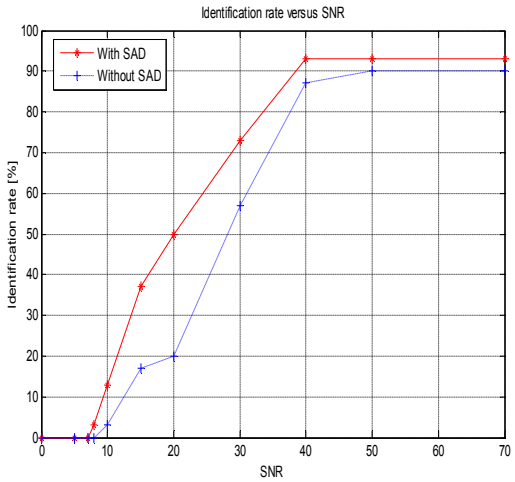


Figure III.20 Taux d'identification avec et sans algorithme de détection d'activité vocale (SAD) vis-à-vis SNR

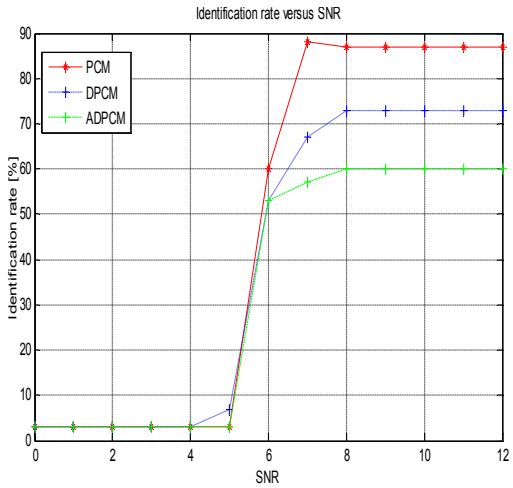


Figure III.21. Effet de PCM, DPCM, et ADPCM sur le taux d'identification vis-à-vis SNR

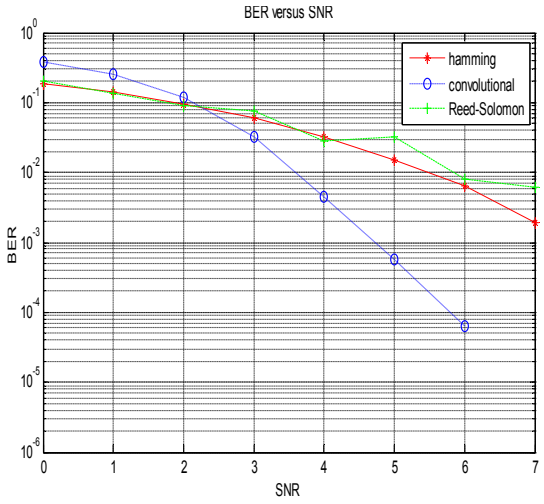


Figure III.22 Étude comparative des techniques de codage: code convolutif, Reed Solomon et Hamming en matière de BER versus SNR

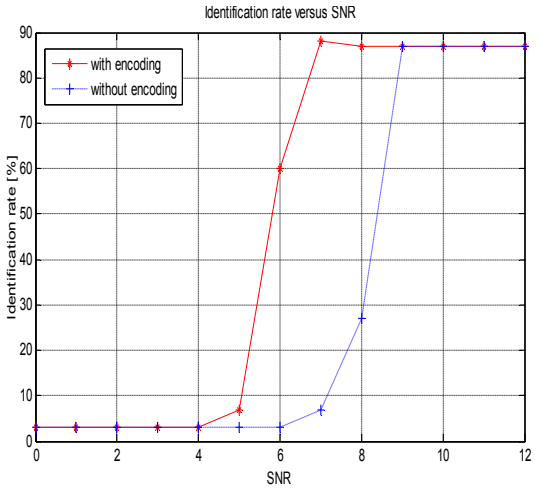


Figure III.22 Taux d'identification avec et sans code convolutif

	RAL	**RAL à travers AWGN**
Identification rate %	93	87

Tableau III.1 Résultats de simulation du taux d'identification à l'aide des signaux parole original et synthétisé après transmission à travers le canal AWGN

	PCM	**DPCM**	**ADPCM**
Temps d'exécution [sec]	100.84	764.24	554.32

Tableau III.2: Temps d'exécution des codecs : PCM, DPCM et ADPCM. (nous avons utilisé un ordinateur portable Intel (R) Core (TM), 103GHz i5-3210M CPU @ 2,5 GHz 2.50GHZ)

III.4 Conclusion

Ce chapitre a présenté une vue d'ensemble sur la reconnaissance du locuteur sur les réseaux mobile, sans fil et internet. On a donné un aperçu sur le signal parole dans les réseaux (Mobile et IP), dont les performances d'un système de reconnaissance automatique du locuteur à distance se dégradent à cause de distorsion du canal, présence de bruits et le codage de signal parole.

Bien que le type de codec (codeur/décodeur) soit un facteur important, dans ce chapitre, nous avons fait une étude comparative des codecs: PCM, DPCM et ADPCM en tenant compte de leurs effets sur les performances de notre système de reconnaissance automatique distant et cela dans un environnement bruité. Etant donné, la détection

d'activité vocale (SAD) effectue la classification parole/silence, nous avons développé un nouvel algorithme d'activité vocale (SAD) qui améliore la capacité mémoire à laquelle l'extraction des paramètres ne s'exerce pas dans les zones classifiées comme des zones de silence ce qui signifie un taux d'identification amélioré.

Notre algorithme proposé SAD basé sur l'énergie et le taux de passage par zéro, donne un contour approprié d'activité de la parole. Notre SAD a fonctionné avec précision dans des environnements de faible SNR (jusqu'à SNR de 5 dB)

Afin d'améliorer le taux de reconnaissance, l'utilisation d'un codage de canal est nécessaire pour rendre le système à distance plus robuste aux erreurs de canal; par conséquent, nous avons choisi le code convolutif. La meilleure performance globale de codecs a été observée pour le codec PCM en termes de taux de reconnaissance et temps d'exécution.

Il est recommandé alors d'utiliser la technique PCM comme codec de la parole dans les systèmes RSR destiné pour les applications VoIP.

Chapitre 4

Développement et évaluation d'un système de RSR

Sommaire

IV.1 Introduction

Les tâches les plus difficiles dans la reconnaissance du locuteur sont l'extraction de paramètres, la détection d'activité vocale (Speech Activity Detection - SAD) et la modélisation du locuteur. Le modèle de mélange gaussien (GMM) est efficace pour les systèmes d'identification du locuteur indépendant du texte qui est le cas de notre système de reconnaissance. Dans ce travail, nous avons utilisé le GMM pour la modélisation des locuteurs. Toutes techniques de RAL commencent par la conversion du signal de parole brute en une séquence de vecteurs des caractéristiques acoustiques porteurs d'informations sur le signal. Cette extraction des caractéristiques est aussi appelée "front end" dans la littérature.

Les vecteurs acoustiques les plus couramment utilisés sont les paramètres MFCC. Cependant, les paramètres MFCC et leurs dérivées à savoir ΔMFCC et ΔΔMFCC sont très utiles dans des conditions propres, mais se détériorent en présence du bruit. En outre, les modèles autorégressifs, sont également importants pour représenter le signal parole.

Dans ce chapitre, nous avons développé un système de reconnaissance du locuteur à distance (RSR) à travers le canal AWGN fondé sur la combinaison des paramètres d'autorégressifs (AR) et MFCC qui s'avère plus robuste en milieu bruité. Afin d'augmenter le taux de reconnaissance, une amélioration d'algorithme de détection d'activité vocale (SAD) vue dans le chapitre précédent est faite en prenant en compte d'estimation de bruit auparavant (Prior SNR estimation) la décision de parole/non-parole.

En d'autres termes, une étude des effets des techniques d'accès multiple nécessaires aux réseaux mobiles et IP à savoir OFDMA, SC-CDMA (L'étalement de spectre en séquence directe) sur notre système RSR est discutée dans ce chapitre.

Enfin, nous nous focalisons sur les dégradations dues à la prise de son (bruits ambiants et variations du canal acoustique), et afin d'améliorer au mieux le taux de reconnaissance en présence du bruit, une étude comparative des techniques de rehaussement (d'élimination de bruit) de signal parole est réalisée avec une application de ces techniques sur notre nouveau système de RAL basé sur MFCCAR et SAD.

VI.2 RSR basée sur une nouvelle approche d'extraction des paramètres (MFCCAR)

La phase d'extraction de paramètres nécessite beaucoup d'attention parce que les performances de la reconnaissance dépendent fortement de cette phase. Les coefficients

cepstreaux (MFCC) sont très utiles pour la reconnaissance du locuteur dans des conditions propres, mais ils se détériorent en présence du bruit. En d'autres termes, les modèles autorégressifs, sont également importants pour représenter le signal de parole et cela justifié par plusieurs travaux dans la littérature (Delima, Charles B. 2002 [112]) et (El Ayadi, M. 2008 [113]). Pour cela, dans notre travail, on propose la combinaison des MFCCs et les paramètres du modèle autorégressif (AR) comme une méthode robuste d'extraction de paramètres dans un environnement bruité.

Dans la phase de prétraitement, nous avons proposé d'améliorer notre algorithme de détection parole /non-parole (SAD) vue dans le chapitre précédent par une estimation de niveau de bruit en avant (prior SNR estimation) et cela par une estimation de la variance de bruit. L'évaluation de l'algorithme SAD a été réalisée en utilisant un corpus de parole bruyante (NOIZEUS) développé dans le laboratoire de Hu et Loizou (Hu,Y et Loizou , P. 2007 [114]), la description de ce corpus a été publiée dans [114]. Pour valider notre approche d'extraction des paramètres du locuteur, on a utilisé la base de données TIMIT [102]. Les quatre premiers énoncés pour chaque locuteur ont été définis comme l'ensemble prévu pour la phase d'apprentissage et un énoncé comme l'ensemble de tests.

L'évaluation de notre nouvelle approche MFCCAR est faite en présence de différentes natures de bruit. Les échantillons de bruit de plusieurs types (bruit gaussien, rose, bleu, violet et bruits blancs) ont été sélectionnés pour valider notre approche. Pour produire artificiellement des signaux parole bruités, les signaux du locuteur ont été ajoutés à chacun des types d'échantillons de bruit à différents niveaux d'SNR.

Le GMM est l'approche la plus courante pour la reconnaissance en mode indépendant du texte et adopté par plusieurs travaux dans la littérature, pour cela dans notre travail, on a utilisé le GMM comme une méthode de modélisation et décision pour la reconnaissance du locuteur.

IV.2.1 Configuration du système proposé

Dans la reconnaissance du locuteur, on distingue deux tâches principales : l'identification et la vérification. Le système que nous allons décrire est classé en tant que système d'identification du locuteur en mode indépendant du texte à distance et qui a été mis en place selon le diagramme représenté par la figure IV.1.

Dans cette section, nous considérons le système de reconnaissance du locuteur à travers un canal de transmission AWGN. Tout système de reconnaissance se compose de trois parties : la phase d'apprentissage, la phase de test et la phase de décision d'y compris

l'identification et la vérification (accepter/rejeter). Dans le système proposé, les étapes de tests et d'apprentissage fondés sur deux blocs principaux, prétraitement où notre algorithme de détection d'activité vocale est utilisé et l'extraction des trais du locuteur où nous avons combiné les paramètres MFCC et vecteurs AR. Le système que nous avons utilisé a été établi selon le schéma suivant dans Figure IV.1.

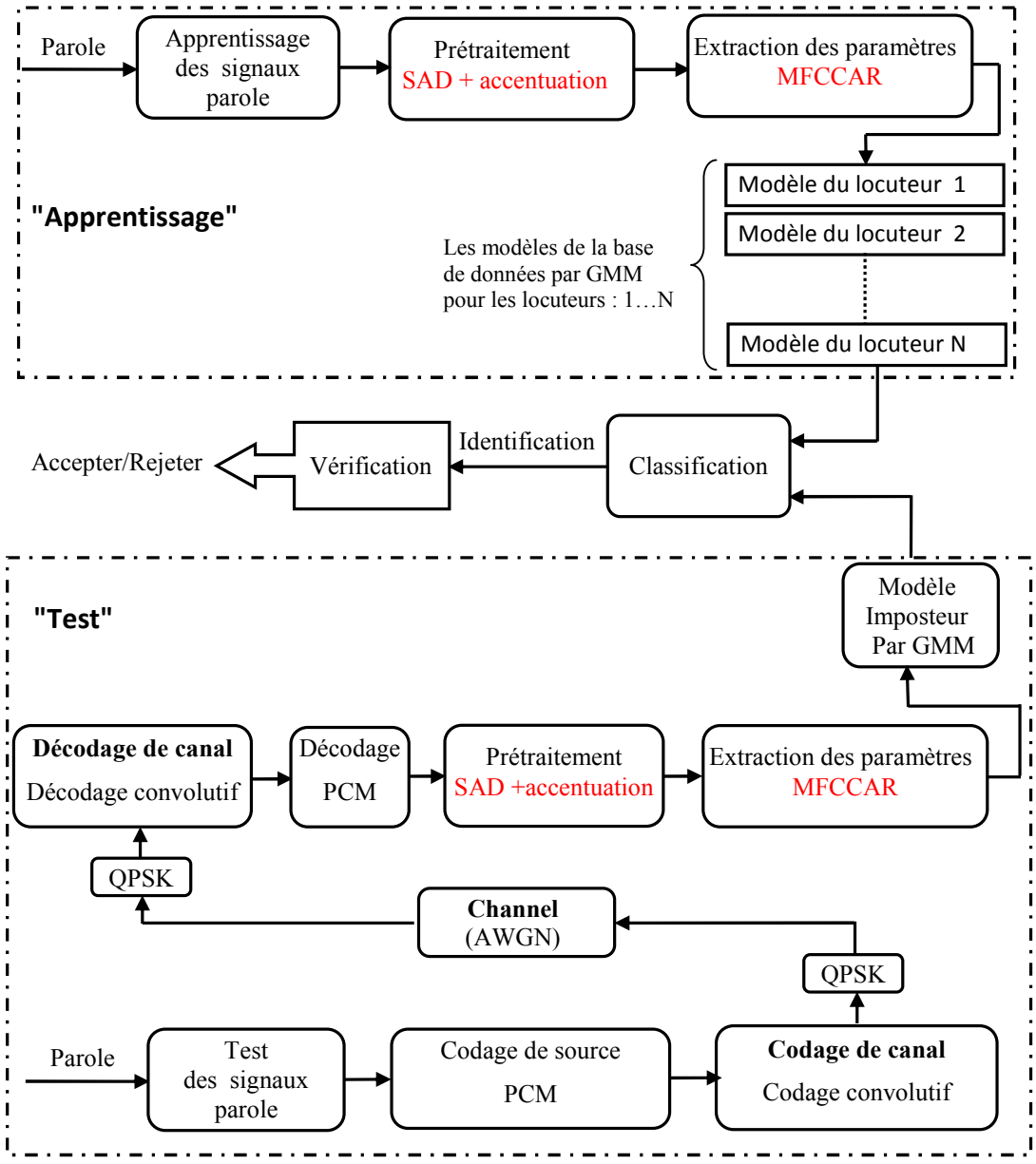


Figure IV.1 Schéma du système RSR proposé basé sur MFCCAR.

IV.2.1.1 Prétraitement

Cette étape est essentielle puisqu'il s'agit d'extraire d'un signal parole, l'information appropriée relative à la tâche de classification, elle est constituée de deux éléments la

préaccentuation, dont chaque signal est accentué par un facteur de 0.97 (équation I.10) et la détection parole/non-parole (les zones de silences sont enlevées par notre algorithme-SAD).

IV.2.2 Technique proposée d'extraction des paramètres (MFCCAR)

L'objectif est d'utiliser des paramètres appropriés pour la reconnaissance du locuteur. Pour cela, dans ce travail, nous avons combiné les paramètres MFCC avec les coefficients du modèle autorégressif (AR). Le nombre des coefficients est 64 (32 MFCC et 32 AR).

Le signal acoustique contient différents types d'informations à propos du locuteur. Les MFCC sont utilisés en reconnaissance de la parole et en identification du locuteur car ces paramètres sont bien adaptés au signal parole et largement utilisés dans la littérature. Ils sont issus de l'hypothèse que le signal parole est le résultat de la convolution entre un filtre qui exprime le conduit vocal et une excitation des cordes vocales (Figure III.4).

Les MFCC permettent une déconvolution entre la source des sons produits (caractéristiques du locuteur) et le conduit vocal (couplé ou non au conduit nasal).

L'échelle non linéaire de Mel (équation I.5) est connue pour rendre compte de la perception humaine qui se compose de banc de filtres. Chaque filtre a de formes triangulaires. Les bancs de filtre triangulaires dans l'échelle de Mel sont espacés uniformément. Dans notre travail, on a limité le nombre de filtres triangulaire à 24 filtres.

Dans cette section, on propose de combiner les coefficients MFCC avec les coefficients autorégressifs AR de filtre tous pôles, les coefficients AR sont ceux du dénominateur du filtre tous pôles dont la transmittance est l'enveloppe spectrale du signal (équation I.5) qui caractérise le conduit vocale. La figure IV.2 représente la procédure de combinaison entre les coefficients MFCC et les paramètres AR dans un seul vecteur MFCCAR, dont on peut résumer la procédure comme suit :

1- fenêtrage de signal parole par des fenêtres de 20ms avec un chevauchement de 10ms (50%).
2- élimination des zones de silence (par notre algorithme SAD).
3- extraction des paramètres MFCC (32 paramètres) et AR (32 paramètres) pour chaque trame dont le nombre de filtres triangulaires utilisés par MFCC est de 24 filtres pour chaque trame.
4- rassembler les paramètres MFCC de toutes les trames dans une seule matrice (MFCC_globale).

5- rassembler les paramètres AR de toutes les trames dans une seule matrice (AR_globale).
6- on résulte le vecteur MFCCAR qui se compose de MFCC_globale et AR_globale MFCCAR= [MFCC_globale AR_globale]. (Voir figure IV.2).

- Coefficients MFCC:

Pour une séquence $s(n)$ qui caractérise un signal parole:

$$s = \begin{bmatrix} s_0 & s_1 & \ldots\ldots & s_{N-1} \end{bmatrix}^T \quad \text{(IV.1)}$$

Où N: représente le nombre d'échantillons de signal parole.

Les coefficients MFCC de signal parole en entier peuvent être représentés par:

$$c = \begin{bmatrix} c_0 & c_1 & \ldots\ldots & c_M \end{bmatrix}^T \quad \text{(IV.2)}$$

Où: M représente le nombre de trames, tandis que $c_0 \quad c_1 \quad \ldots\ldots \quad c_M$, représentent les MFCCs pour toutes les trames représentant le signal parole $0,1.....M$. c_0 represente les MFCCs pour la trame noméro "0" qu'on peut exprimer par:

$$c_0 = \begin{bmatrix} cc_{00} & cc_{01} & \ldots\ldots & cc_{0L} \end{bmatrix}^T \quad \text{(IV.3)}$$

L : représente le nombre des coefficients MFCC qu'on va adopter (Dans notre travail nous avons considéré $L = 32$). De les équations IV.2 et IV.3, on aura les coefficients MFCC globale pour le signal parole considéré par la matrice suivante:

$$c = \begin{bmatrix} cc_{00} & cc_{10} & cc_{20} & \ldots\ldots\ldots cc_{M0} \\ cc_{01} & cc_{11} & cc_{21} & \ldots\ldots\ldots cc_{M1} \\ & & & \\ cc_{0L} & cc_{1L} & cc_{2L} & \ldots\ldots\ldots cc_{ML} \end{bmatrix} \quad \text{(IV.4)}$$

- Les vecteurs Autoregressive (AR):

L'objectif essentiel de la modélisation AR d'un signal parole est de permettre sa description par un ensemble de paramètres (voir I.6.1.1).

Pour une séquence $s(n)$ représentant le signal parole (equation IV.1), les coefficients AR pour le signal en entier peuvent être représentés par:

$$A = \begin{bmatrix} A_0 & A_1 & ..... & A_M \end{bmatrix}^T \quad \text{(IV.5)}$$

dont M: représente le nombre de trames. $A_0 \quad A_1 \quad ..... \quad A_M$ represente les coefficients des vecteurs AR pour les trames: $0,1.....M$. Pour la trame numéro "0" les coefficients AR peuvent etre représentés par la matrice A_0 représeté par:

$$A_0 = \begin{bmatrix} AA_{00} & AA_{01} & ..... & AA_{0p} \end{bmatrix}^T \quad \text{(IV.6)}$$

dont $p = 32$ est le nombre de parameters AR à considerer.

De l'équation IV.5 et IV.6 on aura la matrice globale qui représente les coefficients AR du signal parole en entier:

$$A = \begin{bmatrix} AA_{00} & AA_{10} & AA_{20} & .....AA_{M0} \\ AA_{01} & AA_{11} & AA_{21} & .....AA_{M1} \\ & & & \\ AA_{0L} & AA_{1L} & AA_{2L} & .....AA_{MP} \end{bmatrix} \quad \text{(IV.7)}$$

- Les coefficients MFCCAR:

De l'équation IV.4 et IV.7 on établit les coefficients MFCCAR qui combinent les parameters MFCC et vecteurs AR (L=p=32):

$$MFCCAR = \begin{bmatrix} cc_{00} & cc_{10} & cc_{20} & .....cc_{M0} & AA_{00} & AA_{10} & AA_{20} & .....AA_{M0} \\ cc_{01} & cc_{11} & cc_{21} & .....cc_{M1} & AA_{01} & AA_{11} & AA_{21} & .....AA_{M1} \\ & & & & & & & \\ cc_{0L} & cc_{1L} & cc_{2L} & .....cc_{ML} & AA_{0L} & AA_{1L} & AA_{2L} & .....AA_{MP} \end{bmatrix} \quad \text{IV.8)}$$

La figure IV.2 représente la procédure d'extraction des paramètres MFCCAR pour un signal parole soumis au fenêtrage, dont Cinque trames représentent l'activité de signal parole (on a développé un algorithme qui détecte les trames paroles/non-parole – voir IV.2.6). L'extraction des MFCCs et ARs sont faits pour chaque trame qui représente la parole.

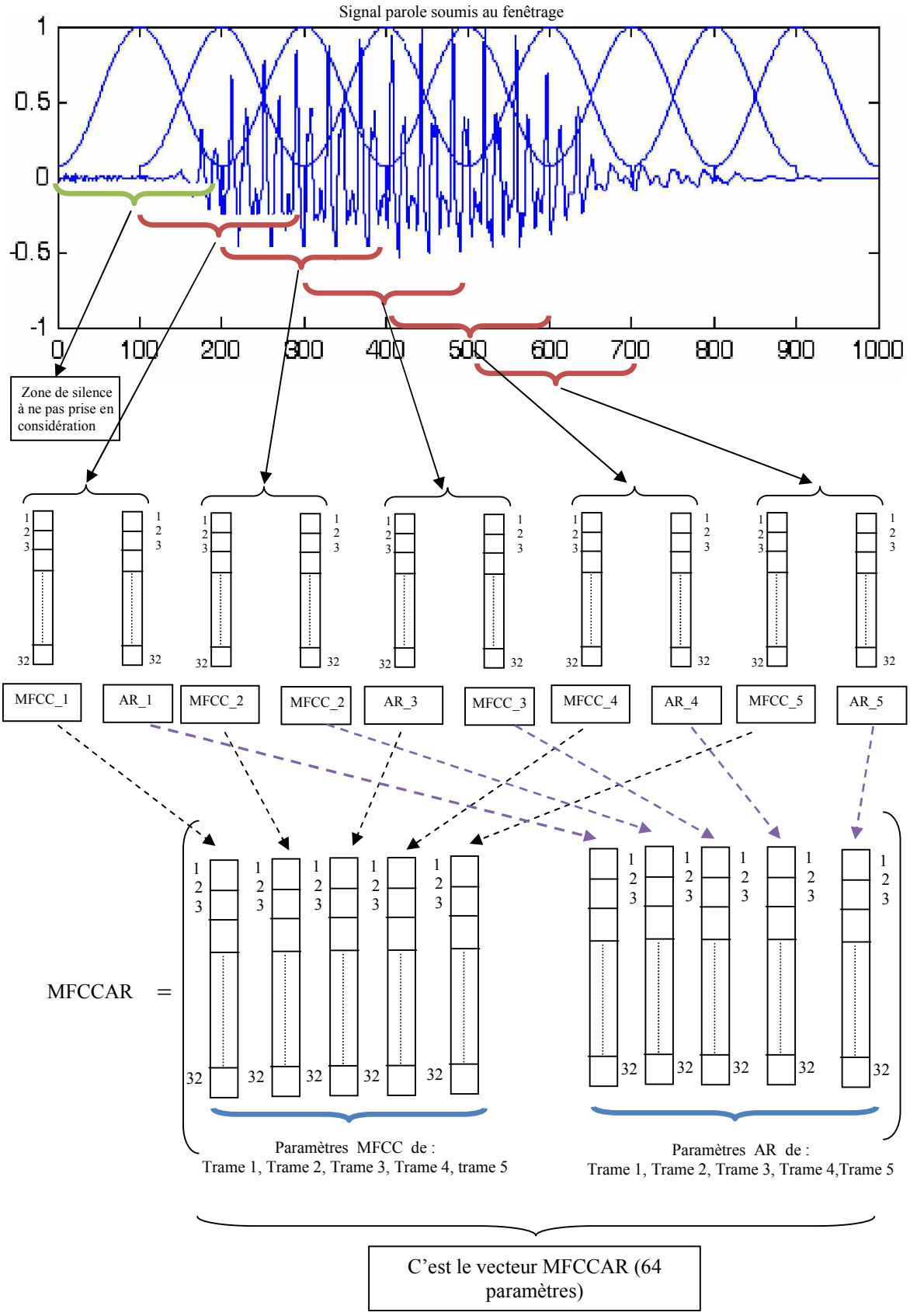


Figure IV.2 Procédure d'extraction des paramètres MFCCAR.
(« MFCC_1, AR_1 », « MFCC_2, AR_2 », « MFCC_3, AR_3 », «MFCC_4, AR_4 » et «MFCC_5, AR_5 » sont les paramètres des trames : trame1, trame 2, trame 3, trame 4, trame 5 respectivement).

IV.2.3 Modélisation des locuteurs par GMM

Dans notre travail, le modèle génératif utilisé est le modèle de mélange gaussien (GMM), la modélisation par GMM fait partie de la phase d'apprentissage. C'est le processus de génération de modèles spécifiques à chaque locuteur avec les données recueillies. Il s'agit, d'estimer les paramètres d'un modèle GMM du locuteur par la méthode du maximum de vraisemblance (MV) (voir chapitre 1).

Par conséquent, il fournit, une représentation statistique du locuteur produit des sons. Il fournit, une représentation statistique sur comment le locuteur produisait les sons. Densité de mélange gaussien est indiquée pour fournir une approximation lisse à la distribution sous-jacente d'échantillon à long terme des observations tirées des énoncés par un locuteur donné.

Le système a été formé à l'aide de la base de données TIMIT [16], dont nous avons choisi 200 intervenants de différentes régions. En outre, à l'étape d'apprentissage, nous avons utilisé quatre énoncés pour chaque locuteur. Le signal parole est passé à travers la phase de prétraitement (préaccentuation + SAD), par la suite soixante-quatre coefficients, sont extraits (32 MFCC et 32 paramètres du modèle autorégressif AR) et la caractérisation des modèles GMM sont formées.

IV.2.4 Phase de test

Dans cette étape, le signal parole codé par PCM, leurs coefficients sont convertis en une séquence binaire. Le code convolutif (utilisé comme un code correcteur d'erreur - FEC) et la modulation QPSK sont utilisées dans notre système de reconnaissance à travers le canal AWGN.

À l'issue du canal de transmission AWGN, les données binaires sont reconverties en un signal parole synthétisé. Enfin, 32 coefficients cepstraux (MFCC) et 32 paramètres (AR) sont extraits (à partir des fichiers paroles synthétisés) et le modèle GMM a été formé.

IV.2.5 Phase de décision (Identification, vérification)

C'est la tâche de calculer les scores correspondants entre les vecteurs des paramètres d'un imposteur modélisé par (GMM) arrivés de la phase de test et les modèles de la base de données (modélisé par GMM) [105]. La phase de décision à savoir l'identification et la vérification est effectuée par le modèle GMM (voir I.6.2.4).

IV.2.5.1 Identification du locuteur

Soit un groupe de ξ locuteurs, représentés par les modèles GMM : $\lambda_1, \lambda_2, \ldots, \lambda_\xi$. L'objectif de la phase d'identification est de trouver, à partir d'une séquence observée X, le modèle qui est à la probabilité a posteriori maximale [43] :

$$\hat{S} = \arg\max_{1 \le s \le \xi} p(\lambda_s / X) \qquad \text{(IV.9)}$$

Ce qui donne, d'après la loi de bayes [43] :

$$\hat{S} = \arg\max_{1 \le s \le \xi} \frac{p(X/\lambda_s)}{p(X)} p(\lambda_s) \qquad \text{(IV.10)}$$

Supposant l'équiprobabilité d'apparition des locuteurs $(p(\lambda_s)=1/\xi)$ et que la probabilité $p(X)$ d'apparition d'une séquence X est la même pour tous les locuteurs, la loi de classification devient [43]:

$$\hat{S} = \arg\max_{1 \le s \le \xi} p(X/\lambda_s) \qquad \text{(IV.11)}$$

En utilisant le logarithme et l'indépendance entre les observations, le système d'identification calcule le score suivant [43]:

$$\hat{S} = \arg\max_{1 \le s \le \xi} \sum_{n=1}^{N} p(X/\lambda_s) \qquad \text{(IV.12)}$$

Où $p(x_n/\lambda_s)$ est donné par l'équation (III.1).

En identification en ensemble fermé, les performances se mesurent en termes de taux d'identification incorrecte I_i et se dégradent considérablement si le nombre de locuteurs ξ augmente.

IV.2.5.2 Vérification du locuteur

La stratégie de décision en vérification du locuteur a fait l'objet de plusieurs travaux de recherche. Dans ce paragraphe, nous présenterons le test d'hypothèse utilisé en

vérification du locuteur et différentes approches d'estimation du modèle du rejet. Ensuite, nous introduirons quelques techniques de normalisation de scores.

En vérification du locuteur, nous utilisons souvent le test d'hypothèses suivant [1]:

- H_0 : le segment de parole X a été prononcé par le locuteur λ.
- H_1: le segment de parole X a été prononcé par un imposteur.

On calcule le rapport de vraisemblance LR (Likelihood Ratio) donné par [1]:

$$LR = \frac{P(X/H_0)}{P(X/H_1)} \underset{>}{\overset{<}{\theta}} \quad \begin{cases} on\ rejete & H_1 \\ on\ accepte & H_0 \end{cases} \tag{IV.13}$$

Où θ est un seuil qui dépend du modèle du locuteur λ. L'hypothèse H_0 correspond au modèle du locuteur λ et l'hypothèse H_1 au modèle de rejet [1].

Généralement, l'hypothèse H_1 correspond à un modèle de rejet indépendant du locuteur appelé modèle du monde λ_{UMB} et l'on évalue plutôt le LLR (log Likelihood Ratio) soit [1] :

$$LLR = \Lambda = \log p(X/\lambda) - \log p(X/\lambda_{UMB}) \tag{IV.14}$$

Où : Le LLR est ensuite comparé à *un seuil*.

Cependant, il est relativement facile d'estimer le modèle du locuteur λ mais il n'en est pas de même pour le modèle du rejet. Ce dernier correspond souvent au modèle du monde [20]. Le modèle de rejet peut aussi correspondre aux cohortes de locuteurs. Cet ensemble peut être spécifique d'un locuteur λ.

Par ailleurs, la normalisation de scores a été introduite, dont, on a utilisé la Z-norme (pZero Normalization) qui consiste à ramener les scores à espace de comparabilité commun [20]:

$$S_{Z-norm} = \frac{\log p(X/\lambda) - \mu_I}{\sigma_I} \tag{IV.15}$$

Où les paramètres de cette normalisation μ_I et σ_I sont estimés, off-line, à partir d'un autre corpus de données. Ils représentent la moyenne et la variance des scores du corpus sur le modèle du locuteur [1]. Dans notre travail, on a utilisé Z-norme pour normaliser les scores (afin d'améliorer la vérification).

Le score de vérification

La décision de vérification s'effectue, en utilisant une valeur estimée empiriquement en comparant le LLR à un seuil ($\theta = -2,5129.10^{3}$ estimée en utilisant la base de données TIMIT), déterminé empiriquement.

IV.2.5.3 Evaluation en vérification du locuteur

Dans la littérature la comparaison des performances de différents systèmes de reconnaissance est effectuée en utilisant la courbe DET, le taux d'égal érreur (EER) et la fonction de coût de décision (DCF) [1]. La courbe DET représente l'évolution des deux types d'erreurs (les faux rejets *FR* et les fausses acceptions *FA* (voir I.7) en fonction du seuil θ. L'EER correspond au point où le taux de FA est égal au taux de FR, la fonction DCF introduit une fonction de coût. Plus l'EER est faible, plus le système est performant ; toutes ces mesures sont calculées globalement sur un grand nombre de tests cible et imposteur. En fonction du type d'application souhaitée, le seuil de vérification peut être choisi pour minimiser le taux de fausses acceptations : application de sécurité, ou minimiser le taux de faux rejets pour augmenter l'ergonomie d'utilisation. Il n'est pas possible de minimiser conjointement ces deux taux. La figure IV.3 illustre l'évolution des taux FA et FR, en représentant l'EER.

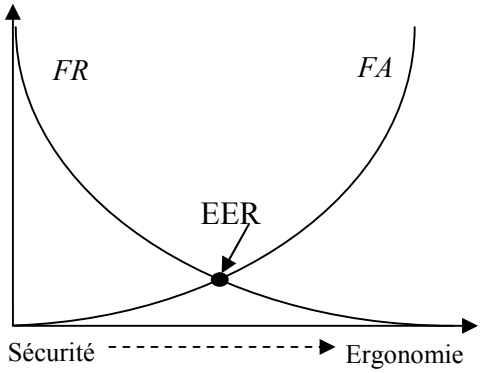


Figure IV.3 Evolution des taux FA et FR.

Point de fonctionnement

Pour comparer les systèmes de RAL deux points de fonctionnement sont extraits pour caractériser plus simplement ces courbes. Le taux d'erreurs égale ou EER (Equal Error Rate) défini comme le point de fonctionnement où FA = FR. À ce point de fonctionnement aucune priorité n'est donnée à la minimisation des FA ou de FR. Cette mesure est très utilisée pour comparer les performances des systèmes de RAL.

Pour introduire une pondération pour chacun de ces taux, en fonction du contexte applicatif, une fonction de coût de détection (DCF, Decision Cost Function) peut être appliquée. Cette DCF s'exprime sous la forme [115] :

$$DCF = C_{FA}\tau_{FA}P_{false} + C_{FR}\tau_{FR}P_{true} \quad \text{(IV.16)}$$

où: - τ_{FA} est le taux de fausses acceptations.

- τ_{FR} est le taux de faux rejets.
- C_{FA} est le coût associé à une fausse acceptation.
- C_{FR} est le coût associé à un faux rejet.
- P_{true} est la probabilité a priori d'un accès client.
- P_{false} est la probabilité d'imposture.

Une autre mesure, nommée HTER (Half Total Error Rate), est définie comme la distribution du taux d'erreur moyen pour chaque seuil de décision (Bengio et al. 2004 [115]).

$$HTER = \frac{1}{2}\left(FA + FR\right) \quad \text{(IV.17)}$$

L'évaluation de notre système a été réalisée en utilisant cette dernière formule HTER en prenant un seuil de vérification fixe : $\theta = -2,5129.10^{3}$

IV.2.6 Algorithme de détection parole/non-parole

La procédure de détection parole/non-parole est résumée comme suit:

1- segmenter le signal parole en entier en trames de 8 ms avec une fenêtre rectangulaire et sans chevauchement.
2- calcul de l'énergie E[m] et le taux de passage par zéro PPZ [m] pour chaque trame
4- calculer le rapport EZR = E [m] / ZCR [m] pour chaque trame.
3- calculer le maximum et le minimum d'EZR.
4- Estimation de seuil de decision parole/non-parole suivant l'equation (III.23 :
$seuil = \min\left(EZR\right) + \alpha \times \left[DELTA\right]$).

Le seuil estimé dépend de paramètre "α", la valeur de α résulte de niveau de bruit, pour cela dans ce chapitre on propose d'estimer le bruit à fin d'éstimer les valeures de α, qui correspondent aux meilleurs taux d'identifications. À ce fin une analyse est faite, le tableau 1 présente les taux d'identifications qui correspondent aux meilleures valeurs de α en terme de meilleur taux d'identification vis-à-vis SNR (10, 20, 30, 50 [dB]). Ces résultats montrent que, pour avoir un meilleur taux d'identification, nous augmentons la valeur de α si le niveau d'SNR aumente. Par exemple si SNR = 30dB, un meilleur taux d'identification correspond à α=0,3.

IV.2.6.1 Mésure d'SNR par estimation de la variance de bruit

Notre algorithme SAD nécessite une connaissance d'SNR pour estimer le seuil. Il y a deux possibilités pour le calcul d'SNR. La première consiste à estimer le rapport de la puissance du signal et la variance du bruit directement, dont la variance du bruit est une mesure de la dispersion statistique de l'amplitude de bruit d'un signal reçu; l'autre est pour obtenir l'estimation de puissance de signal et l'éstimation de la variance du bruit. L'estimation d'SNR soit par [116]:

$$\mathrm{SNR}_{dB} = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_v^2}\right) \qquad \text{(IV.18)}$$

Où σ_x^2 est la valeur de la variance du signal et σ_v^2 est la valeur de la variance du bruit. Dans cette section, nous décrivons la méthode d'estimation de la variance de bruit du segment donné du signal bruité. En présence d'un bruit blanc gaussien additif $v(n)$, le signal observé $y(n)$ peut être écrit comme:

$$y(n) = x(n) + v(n) \qquad \text{(IV.19)}$$

Où $x(n)$ est le signal non contaminée, $v(n)$ est le bruit blanc de moyenne nulle et de variance σ_v^2. Le but est d'estimer la variance de bruit σ_v^2, à partir du signal bruité observé $y(n)$. Afin de résoudre ce problème, nous supposons que le signal non contaminée $x(n)$ suit le modèle AR avec p-ième ordre (équation I.4). Les paramètres AR satisfont l'ensemble des équations de Yule-Walker suivante [117]:

$$\sum_{k=1}^{p} a_k R_x(|i-k|) = -R_x(i) \qquad i > 0 \qquad \text{(IV.20)}$$

où $R_x(i)$ sont les coefficients d'autocorrélation du signal $x(n)$ non contaminé. Etant donné que le bruit additif $v(n)$ est blanc, les coefficients d'auto-corrélation $R_x(i)$ du signal non contaminé $x(n)$ sont liés à des coefficients d'auto-corrélation $R_y(i)$ du signal bruité $y(n)$ comme suit [117]:

$$R_x(0) = R_y(0) - \sigma_v^2 \tag{IV.21}$$

et:

$$R_x(i) = R_y(i) \tag{IV.22}$$

Nous avons trois étapes pour estimer la variance de bruit σ_v^2. Ces étapes sont décrites ci-dessous [116] :

Étape 1: de l'observé de signal bruité $y(n)$, calculer les estimations impartiales des coefficients d'autocorrélation $R_y(i)$, i = 0,1.., p + q. Où q > p.

Etape 2: calculer l'estimation des moindres carrés des coefficients AR par la méthode de Cadzow [118] du q (> p) d'ordre supérieur équations de Yule-Walker [i = p + 1, p + 2, ..., p + q].

Étape 3: utilisez les coefficients AR obtenus à partir de l'étape 2 et calculer l'estimation des moindres carrés de la variance de bruit de l'ensemble des surdéterminés de 'p' d'ordre faible équations de Yule-Walker [i = 1, 2….p]. Cela est donné par [118]:

$$\hat{\sigma}_v^2 = \left[\sum_{i=1}^{p} a_i \left\{ \hat{R}_x \mathrm{x}(i) + \sum_{k=1}^{p} a_k \hat{R}_y \left(|i-k| \right) \right\} \right] \Big/ \sum_{i=1}^{p} a_i^2 \tag{IV.23}$$

Après estimation de la variance du bruit, nous estimons la valeur de la variance du signal et calculons la variance du signal bruité (signal parole + AWGN):

$$\hat{\sigma}_x^2 = \hat{\sigma}_y^2 - \hat{\sigma}_v^2 \tag{IV.24}$$

De l'équation IV.18 on aura :

$$\mathrm{SNR}_{dB} = 10\log_{10}\left(\hat{\sigma}_x^2\right) - 10\log_{10}\left(\hat{\sigma}_v^2\right) \tag{IV.25}$$

L'algorithme SAD proposé est décrit par le schéma de la figure IV.4.

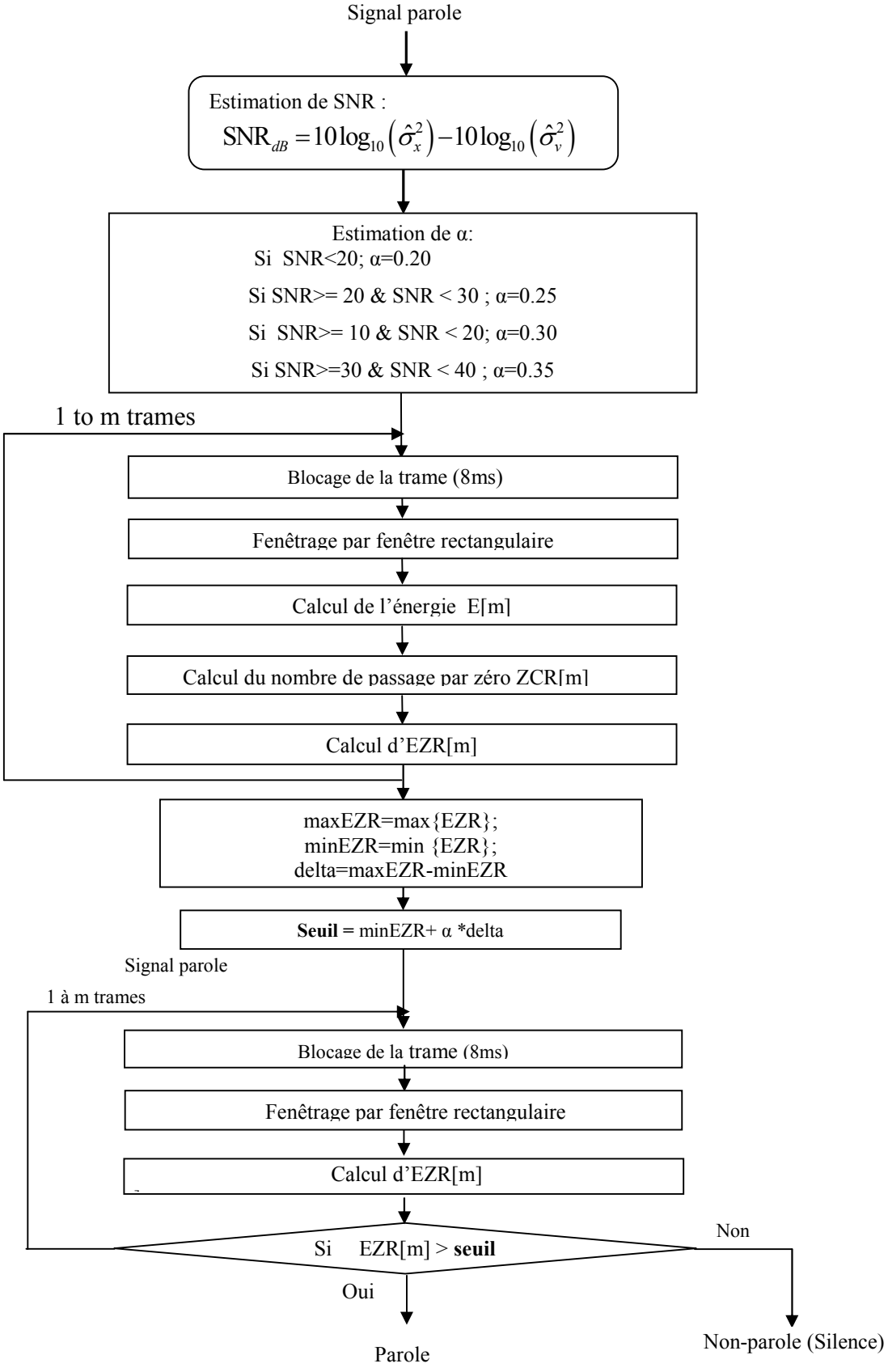


Figure IV.4 Algorithme proposé de détection parole/non-parole.

IV.3. Comparant CDMA et OFDMA sur la performance de notre système RSR

Dans cette section, on présente deux techniques d'accès multiples OFDMA et CDMA et étudiant leurs effets sur notre système de reconnaissance du locuteur à travers le canal de Rayleigh en matière des taux de reconnaissance et d'erreur binaire (BER).

IV.3.1 OFDMA (Orthogonal Frequency Division Multiple Access)

L'OFDMA (*Orthogonal Frequency Division Multiple Access*) est une technique de multiplexage et de codage des données utilisée principalement dans les réseaux de téléphonie mobile de 4^{ième} génération. Ce codage radio associe les multiplexages en fréquence et en temps; c'est-à-dire les modes « accès multiple par répartition en fréquence » (AMRF ou en anglais *FDMA*) et « Accès multiple à répartition dans le temps » (AMRT ou en anglais *TDMA*) [121]. Il est notamment utilisé dans les réseaux de téléphonie mobile 4G-LTE, LTE-Advanced et WiMAX (IEEE 802.16e). L'OFDMA ou l'une de ses variantes (SC-FDMA) sont aussi utilisées dans d'autres systèmes de radiocommunication, telles les versions récentes des normes de réseaux locaux sans fil WIFI (IEEE 802.11, IEEE 802.22 ….) ainsi que par certaines normes de télévision numérique. Le codage OFDMA consiste en un codage et une modulation numérique d'un ou plusieurs signaux binaires pour les transformer en échantillons numériques destinés à être émis sur une (ou plusieurs) antennes radio ; réciproquement, en réception, le signal radio reçoit un traitement inverse [121], (OFDM avec plus de détail est en annexe B).

IV.3.2 CDMA (Code division multiple Access)

C'est l'accès multiple par répartition en code, est un système de codage des transmissions, utilisant la technique d'étalement de spectre. Il permet à plusieurs liaisons numériques d'utiliser simultanément la même fréquence porteuse. Ce système est appliqué dans les réseaux de téléphonie mobile (3G UMTS) dans le segment d'accès radio, le principe est l'utilisation simultanée de plusieurs codes.

Le DS-CDMA (*Direct-Sequence Code-Division Multiple-Access*), ou CDMA à séquence directe, c'est la technique d'étalement la plus répandue dans les systèmes de radiocommunication mobile. D'autres applications utilisent l'étalement de spectre : GPS (Global Positioning System), certains réseaux WLAN (Wireless, IEEE 802.11) [122]. La DS-CDMA est un multiplexage à étalement de spectre, dont, chaque utilisateur émet sur

toute la largeur de bande du canal de communications une clé (ou code) propre à chaque utilisateur et permet de coder et décoder les messages [121, 122].

IV.3.2.1 Principes du DS-CDMA

Le CDMA réalise un étalement de spectre selon la méthode de répartition par séquence directe (Direct Sequence). Pour cela, chaque bit de l'utilisateur à transmettre est multiplié (OU exclusif) par un code pseudo-aléatoire PN (Pseudo random Noise code) propre à cet utilisateur. La séquence du code (constituée de N éléments appelés "chips") est unique pour cet utilisateur en question, et constitue la clé de codage. Cette dernière est conservée si le symbole de données est égal à 1, sinon elle est inversée. La longueur L du code est appelée facteur d'étalement SF (Spreading Factor). Si chacun des symboles a une durée Tb, on a 1 chip toutes les Tb/N secondes. Le nouveau signal modulé a un débit N fois plus grand que le signal initialement envoyé par l'usager et utilisera donc une bande de fréquences N fois plus étendue. Un grand nombre de techniques existe pour construire des codes ayant de bonnes propriétés (PN séquence code, les codes de Gold…) [122, 123, 124, 125, 126, 127]. (Le DS-CDMA avec plus de détails est en annexe C).

Le schéma général de transmission de signal parole par OFDMA et DS-CDMA est représenté par la figure IV.5. Après avoir transmis le signal parole et extraire les paramètres MFCCAR du signal reconstitué, on passe à l'étape de reconnaissance.

Les résultats de comparaisons d'OFDMA et DS-CDMA en vue de leur impact sur notre système de RSR sont représentés par le tableau IV.7.

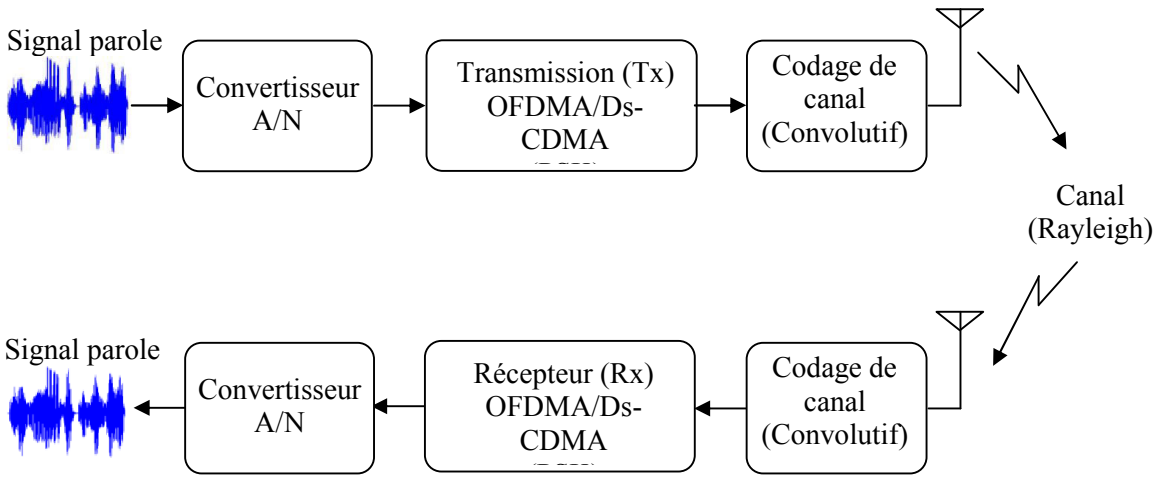


Figure IV.5 Schéma général de transmission de signal parole par OFDMA et DS-CDMA.

IV.4 Étude de différentes techniques d'élimination de bruit additif au signal parole

Les dégradations de la parole imposée par divers réseaux téléphoniques ont été prouvées d'avoir des effets importants sur la performance des systèmes de RAL. Le rehaussement de la parole améliore la qualité et l'intelligibilité de la communication vocale pour une gamme d'application, notamment les téléphones mobiles, les systèmes de téléconférence, les prothèses auditives, des codeurs de la parole et la reconnaissance automatique du locuteur/parole. Un signal peut être corrompu par le bruit dans diverses situations, comme les trains, les voitures, l'aéroport, bavardage, usines, rue....etc. On présente un aperçu des méthodes monocanaux de rehaussement ainsi leurs effets sur notre system d'identification du locuteur basé sur MFCCAR. Le problème de rehaussement ou de débruitage consiste à enlever le bruit à partir du signal corrompu sans l'altérer. Par conséquent, l'étude comparative était en termes de qualité de la parole. La qualité de la parole reconstruite est mesurée avec le PESQ (Perceptual Evaluation of Speech Quality) [81]. Les méthodes que nous avons évaluées sont :

- Tracking Of Non-Stationary Noise Based On Data-Driven Recursive Noise Power Estimation [128], cette méthode est proposée par (Erkelens et al. 2008) [128]. Les auteurs considèrent estimation de la variance du bruit spectral de signaux de parole contaminés par les sources de bruit fortement non stationnaires. La méthode permet de suivre avec précision les changements rapides du niveau de la puissance de bruit (jusqu'à environ 10 db / s). L'algorithme de rehaussement est basée sur l'erreur quadratique moyenne minimale (MMSE) [80] estimation dans le domaine DFT (transformée de Fourier discrète) du l'amplitude spectrale de la parole. L'estimation de la puissance de bruit par MMSE est de mettre à jour les estimations de spectre de bruit et réduire le risque d'altérer la parole. Les estimations MMSE sont obtenues avec la méthode standard [80] par une multiplication de la puissance de bruit et une fonction de gain spectrale. Cela supprime en plus la contribution de la parole à partir du spectre bruyant, permettant un suivi rapide et précis de l'évolution des niveaux de bruit [128].
- Speech Enhancement Based On A Priori Signal To Noise Estimation [129]. cette méthode a été proposée par (Scalart et al. 1996) [129]. Parce que l'estimation a priori de SNR conduit aux meilleurs résultats subjectifs. Selon ces conclusions, une approche a été développée sur la base de filtre de Wiener par le suivi d'un priori SNR en utilisant la méthode de Décision-Direct (DD), $SNRpost = SNRprior + 1$. Dans cette méthode, il est supposé que: sur la base de

cette relation le filtre de Wiener peut être adapté à un modèle comme le modèle de Ephraims dans [130]. Modèle dans lequel nous avons une fonction de gain qui est en fonction d'un priori SNR. Priori SNR est suivi en utilisant la méthode de décision directe DD ($SNRpost = SNRprior + 1$).

- Geometric approach to spectral Geometric (GA) [131]. Cette méthode récente a été proposée par Yang Lu, Philipos C. Loizou (2008) [131] qui est une approche géométrique de soustraction spectrale. Yang Lu et al a présenté un algorithme géométrique (GA) pour soustraction spectrale basée sur des principes géométriques [131]. Contrairement à l'algorithme de soustraction spectrale de puissance classique qui suppose que les termes croisés portants sur la différence de phase entre le signal et le bruit sont égaux à zéro, l'algorithme ne fait pas de telles hypothèses. Cela a été confirmé par l'analyse d'erreur qui indique que, bien qu'il soit sûr d'ignorer les termes croisés lorsque le SNR spectrale et soit très élevé ou très faible, il n'est pas sûr de le faire lorsque le SNR spectrale tombe proche de 0 dB [131]. Un procédé pour incorporer les termes croisés impliquant des différences de phase entre les signaux de bruit (et propre) et le bruit a été proposé [131]. L'analyse des courbes de suppression de l'algorithme GA a indiqué qu'elle possède des propriétés similaires à celles de l'algorithme MMSE traditionnel (Ephraim et al. 1985) [130]. L'évaluation objective de l'algorithme GA a montré qu'il nettement mieux que l'algorithme de soustraction spectrale traditionnelle dans toutes les conditions [131].
- Harmonic Regeneration Noise Reduction (HRNR) [132]. Cette méthode a été proposée par Cyril Plapou, Claude Marro, et Pascal Scalart (2006) [132]. La décision dirigée bien connue (DD) approche limité drastiquement le niveau de bruit musical mais l'estimation d'SNR à priori est biaisée car elle dépend de l'estimation du spectre de la parole dans la trame précédente [132]. Par conséquent, la fonction de gain correspond à la trame précédente au lieu de celle en cours qui dégrade les performances de réduction de bruit [132]. Les auteurs ont proposé une méthode appelée Two-Step Noise Reduction (TSNR) qui résout ce problème tout en conservant les avantages de l'approche DD. L'estimation de la SNR a priori est affinée par une deuxième étape pour éliminer le biais de l'approche DD, éliminant ainsi l'effet de réverbération. Cependant, les techniques de rehaussement classiques à court terme, y compris les TSNR, introduisent une distorsion harmonique dans le signal amélioré en raison de la non-fiabilité des estimateurs pour les petites SNR.

Ceci est principalement dû à la tâche difficile pour estimer la DSP dans les régimes de monocanaux. Pour surmonter ce problème, une méthode appelée harmonique de réduction de la régénération du bruit (HRNR) a été proposée. Une non-linéarité est utilisée pour régénérer les harmoniques dégradées du signal déformé de manière efficace. Une amélioration significative est portée par HRNR par rapport à TSNR grâce à la préservation d'harmoniques [132].

- Phase Spectrum Compensation (PSC) [133]. Cette technique est proposée par (Anthony et al., 2008) [133]. Le spectre de magnitude bruité est recombiné avec un spectre de phase compensé pour une distorsion bruit additive à fin de produire un spectre complexe modifié. Estimations de bruit sont incorporées dans la procédure de compensation de spectre de phase. Lors de la synthèse des composants de basse énergie du spectre complexe modifié annule plus que les composants de haute énergie, réduisant ainsi le bruit de fond [133].
- Speech Enhancement Using a Non causal a Prior SNR Estimator [134]. Cette technique proposée par (I. Cohen, 2004 [134]) basée sur un estimateur d'SNR à priori, et l'erreur quadratique moyenne minimale (MMSE). L'auteur propose un estimateur non causal d'SNR à priori, et un algorithme non causal de rehaussement de la parole correspondant. Contrairement à la DD (Décision directe) estimateur d'Ephraïm et al. [130]. Sur les plateaux de la parole sont mieux conservés, si une nouvelle réduction du bruit soit atteinte. Les résultats expérimentaux montrent que l'estimateur non causal conduit à une amélioration plus élevée dans le SNR segmentaire (segSNR), moins de distorsion de log- spectrale, et bonne PESQ [134].
- Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay [135]. Cette méthode est proposée par (Gerkmann et al. 2012) [135] pour estimer la densité spectrale de puissance de bruit au moyen de l'estimation optimale d'erreur quadratique moyenne minimale (MMSE) [135]. Dans le cas contraire, l'estimateur qui en résulte peut-être interprété comme un détecteur d'activité vocale (VAD) basé sur l'estimateur de puissance de bruit, dont la puissance de bruit est mise à jour uniquement lorsque l'absence de signal parole a été signalé, compensé par une correction biaise [135]. La compensation n'est pas nécessaire lorsque nous remplaçons le VAD par une probabilité de présence de paroles douces (speech présence probability SPP) avec des fixations a priori. Choisir des fixations a priori a également l'avantage de dissocier l'estimateur de

puissance de bruit à partir des étapes ultérieures dans le cadre de rehaussement de la parole, comme l'estimation de la puissance et l'estimation de la parole propre [135]. En outre, l'approche SPP proposée maintient la performance du suivi rapide de bruit de du biais compensé MMSE approche tout en présentant moins surestimation de la puissance spectrale de bruit et une complexité de calcul encore plus bas [135].

IV.5 Résultats et discussion

L'évaluation de la méthode d'extraction des paramètres proposée a été réalisée par des expériences de reconnaissance du locuteur en mode indépendant du texte sur la base de données de TIMIT. Nous avons choisi des locuteurs de différentes régions et désignant les quatre premiers énoncés pour chaque locuteur pour faire l'apprentissage et un énoncé pour faire le test, dont la vérification est procédée par un seuil fixe $\theta = -2,5129.10^3$ (formule IV.5). Les fichiers de la base de données TIMIT sont échantillonnés avec un taux de 16000 échantillons /s, ces fichiers ont été sous échantillonnés à un taux de 8000 échantillons /s. Le signal de la parole est segmenté en trames. Le traitement a été effectué à l'aide de la fenêtre Hamming, la taille de la fenêtre est de 20ms correspond à 160 échantillons avec un chevauchent 10 ms (80 échantillons).

Dans notre travail, le nombre de filtres utilisé est de 24 de forme triangulaire (banc de filtre). Chaque trame est convertie en 32 MFCCs et 32 coefficients autorégressifs. Ces paramètres (MFCCAR) sont utilisés pour former le MGM. Le GMM constitue la base à la fois pour l'apprentissage et les processus de classification. Nous avons fixé un nombre maximal de 100 itérations pour modéliser la voix du locuteur par GMM pour un objectif de réduire la complexité de calcul.

IV.5.1 Démontrer la performance de l'algorithme SAD

L'évaluation de l'algorithme SAD a été effectué en utilisant le corpus des paroles bruités « NOIZEUS » développé dans le laboratoire d'Hu et Loizou [119], la description de ce corpus a été publié dans [81]. La figure IV.6 illustre un signal parole (Signal propre) de la base de données NOIZEUS corpus et son contour d'activité de la parole SAD.

Les figures IV.7, IV.8, IV.9 et IV.10 représentent des exemples des contours d'activité vocale d'un signal parole émergé dans un bruit additif de bavardage (babble) de niveaux différents (15 dB, 10 dB, 5 dB et 0 dB respectivement) à l'aide de l'algorithme de SAD.

Ces figures indiquent que l'algorithme SAD est robuste pour les faibles SNR où le SAD peut décerner les zones de paroles.

IV.5.2 Impact de l'ordre du modèle sur le taux de reconnaissance et le taux d'erreur moyen (HTER)

Dans cette expérience, nous étudions l'impact de l'ordre des modèles (ou le nombre de gaussiennes) sur le taux d'erreur moyen (HTER) et temps d'exécution moyenne (Nous avons utilisé un Lap top: Intel (R) core (TM) i5-3210M CPU, 2.5GHZ 2.50GHZ)sur les système de reconnaissance du locuteur à distance (RSR) en utilisant les coefficients MFCCAR, MFCC, ΔMFCC et PLP dont le cas où on a utilisé 100 locuteurs (4 énoncés) client et 100 imposteur de la base de données TIMIT.

Le tableau IV.2 représente les variations des taux d'erreur moyens HTER, FA et FR en fonction de l'ordre des modèles pour les coefficients, MFCCAR, MFCC, ΔMFCC et PLP. Le tableau IV.2 montre que l'augmentation du nombre de gaussiennes dans la représentation du locuteur apporte une amélioration du taux d'identification mais le temps de calcul augmente considérablement. Le temps d'exécution est moins petit pour le cas de PLP mais plus grand pour MFCCAR. Mais il est intéressant de noter que le choix de l'ordre du modèle dépend de la quantité de données (la longueur du signal parole) d'apprentissage. Choisir un ordre trop peu élevé va nuire la précision du modèle. Ainsi, choisir un trop de composantes engendrera une charge de calcul plus importante. En général, 128 composantes suffisent pour représenter un locuteur dans notre cas de la base de données TIMIT. Concernant le taux d'erreur moyen (HTER) sont comprise entre 5% et 42% pour le cas des systèmes de la reconnaissance basés sur MFCCAR, MFCC, ΔMFCC et PLP en utilisant un seuil de vérification fixe θ = - 2.5129 10^3. Ainsi, Les taux d'identification sont meilleurs pour le cas des systèmes basés sur MFCCAR que les systèmes basés sur MFCC, ΔMFCC et PLP. Pour un nombre de modèles petits, les performances d'identification sont mauvaises dans le cas des coefficients MFCC, mais suffisamment bien pour les coefficients MFCCAR, ΔMFCC et PLP.

IV.5.3 RAL par: MFCCAR, MFCC, ΔMFCC et PLP en présence de différentes natures de bruits (WGN, rose, bleu et violet)

En d'autre terme, on a évalué notre système en présence de différents types de bruits, bruit WGN (bruit blanc gaussien), rose, bleu et violet (n'est à travers le canal de communication) en terme du taux d'identification I_d %. Bruit rose est utilisé pour

remplacer les bruits ambiants dans les expériences liés au son. Il est également utilisé dans les théâtres et les studios où les oreilles humaines doivent évaluer la qualité du son [120]. Nous avons ajouté les différents types de bruit aux signaux parole de la base de données TIMIT (nous avons utilisé 300 signaux du locuteur et l'ordre de GMM = 64). Tous les résultats sont présentés au tableau IV.3. Le tableau IV.3 montre en comparant les techniques d'extractions de paramètres MFCCAR, MFCC, ΔMFCC et PLP que MFCCAR a des taux d'identifications supérieures.

En terme, du taux d'identification moyen, le tableau 4, présente les taux d'identification moyens par MFCCAR, MFCC, ΔMFCC et PLP, dont on remarque que MFCCAR a les plus grand taux d'identifications moyens.

IV.5.4 RAL à travers le canal AWGN par MFCCAR versus MFCC, ΔMFCC et PLP versus SNR.

Dans cette sous-section on compare notre système de reconnaissance du locuteur à travers le canal AWGN basé sur MFCCAR versus MFCC, ΔMFCC et PLP sous condition bruitée (différents niveaux d 'SNR). Ces résultats sont illustrés sur la figure IV.11. D'après ces résultats, on peut conclure que notre approche d'extraction de paramètres MFCCAR fournit des améliorations d'identification du locuteur par rapport aux MFCC et ΔMFCC et PLP à travers le canal AWGN.

IV.5.5 Simulation des effets des techniques OFDMA et DS-CDMA sur RSR

Dans cette sous-section, on étudie l'effet d'OFDMA et DS-CDMA sur notre système de reconnaissance du locuteur à distance (RSR) basé sur MFCCAR et SAD. Les parameters de simulations sont illustrés dans les deux tableaux IV.5 (OFDMA) et IV.6 (CDMA). Nous utilisons donc des fichiers audio originaux et reconstruits après la transmission à travers le canal Rayleigh.

Le tableau IV.7 montre les résultats de simulation du taux de reconnaissance de notre système RSR en utilisant OFDMA et DS-CDMA. Dans nos expériences, nous avons choisi 200 locuteurs d'une façon aleatoire de la base de données TIMIT (avec GMM=64, MFCCAR=64 paramètres). À partir de tableau IV.7, il est facile de constater que le DS-CDMA n'influe pas considérablement sur le taux d'identification par contre le système d' OFDMA influe largement. Il est facile de constater que le DS-CDMA donne un taux d' identification meilleur que le système d'OFDMA et cela à cause que le BER pour le DS-CDMA est mois que celle d'ODFMA. Ainsi, la remarque la plus importante est que le

système de reconnaissance en utilisant le DS-CDMA est plus robuste au bruit contrairement à l'OFDMA et cela justifié par les codes utilisés par le DS-CDMA qui rend le système plus robuste au bruit.

IV.5.6 Comparaison des méthodes de rehaussement de la parole et leurs effets sur notre système de RAL

Sept méthodes de l'état de l'art ont été discutées et évaluées en termes de robustesse contre un bruit réel (bruit de train de banlieue, bavardage, voiture, salle d'exposition, restaurant, rue, le bruit de l'aéroport et de là gare) et des bruits artificiels (roses, bleu, rouge, Violet et bruit blanc gaussien) en utilisant la base de données de TIMIT [102] et le corpus NOIZEUS des signaux parole développée dans le laboratoire de Hu et Loizou (Yi Hu et al. 2008) [119] qui est adaptée à l'évaluation des algorithmes de rehaussement. Il est important d'éclairer que les paramètres de simulations des sept méthodes mentionnées ci-dessus sont les mêmes utilisés par leurs auteurs d'origine.

IV.5.6.1 Comparaison en présence d'un bruit blanc gaussien

Les sept méthodes mentionnées ci-dessus sont évaluées en termes de PESQ et en présence du bruit blanc gaussien (WGN) en utilisant la base de données TIMIT.

La figure IV.12 illustre une comparaison des performances de sept techniques de rehaussement en termes de scores PESQ en présence du bruit blanc pour différents niveaux d' SNR, de -5 à 30 dB par un pas de 5 dB. De cette figure (figure IV.12) on peut conclure que la méthode d'Erkelens et al. (2008) [128] présente le meilleur score de PESQ, tandis que la technique de Gerkmann et al. (2012) [135] comme la seconde en terme de PESQ, par contre la méthode de Scalart et Vieira (1996) [129] la plus mauvaise.

IV.5.6.2 Comparaison en présence des bruits de: bavardage, aéroport, voiture, rue, restaurant et salle d'exposition en utilisant la base de données NOIZEUS.

Les approches de rehaussement de la parole ont également été évaluées en présence de plusieurs bruits réels extraits de la base de données NOIZEUS: bavardage, aéroport, voiture, rue, restaurant et salle d'exposition, sous différents niveaux de bruit (SNR varie entre 0 dB et 15 dB par un pas de 5 dB).

La Figure IV.13 présente une mesure de PESQ pour les sept approches comme mentionnées plus haut, en présence du bruit de bavardage. On remarque que la méthode

« Tracking Of Non-Stationary Noise Based On Data-Driven Recursive Noise Power Estimation » proposée par Erkelens et al. (2008) [128] et l'approche « Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay » de Gerkmann et Richard (2012) [135] ont presque le même PESQ pour SNR plus de 5 dB, mais, pour de faibles niveaux de SNR" (Gerkmann et al. 2012) [135] fournit le meilleur résultat. En outre, la méthode Harmonic Regeneration Noise Reduction (HRNR) [132] proposée par Plapous et al. 2006 [132] est la mauvaise en dessous de 10 dB.

La figure IV.14 représente les mesures de PESQ en présence du bruit d'aéroport. Dont cette figure montre que la méthode d'Erkelens et al. (2008) [128] fournit de meilleurs résultats de PESQ.

La figure IV.15 représente la mesure de PESQ en présence du bruit de voitures. De cette figure, on peut conclure que l'approche d'Erkelens et al. (2008) [128] fournit de meilleurs résultats de PESQ. En outre, la méthode de P. Scalart (1996) [1429] a présenté le mauvais score de PESQ.

La Figure IV.16 représente la mesure de PESQ en présence du bruit de la rue. La technique d'Erkelens et al. (2008) [128] montre le meilleur résultat de PESQ.

La figure IV.17 représente la mesure de PESQ en présence du bruit de restaurant. De cette figure, nous pouvons conclure que les techniques d'Erkelens el al. (2008) [128] et Gerkmann et al. (2012) [135] ont les mêmes résultats de PESQ.

La Figure IV.18 représente mesure de PESQ en présence du bruit de salle d'exposition. De cette figure, nous pouvons conclure que la méthode de Erkelens et el.(2008) [128] était le premier en matière de PESQ.

Tableau IV.8 représente les mesures moyennes de PESQ pour les méthodes mentionnées précédemment pour la parole contaminée par le bavardage, l'aéroport, la voiture, la rue et le bruit du restaurant. À partir de ce tableau, il est aisément de conclure que la méthode Erkelens et al. (2008) [128] a le meilleur score de PESQ.

IV.5.6.3 Résultat pour le cas de signal corrompu par les bruits colorés

Nous évaluons les différentes techniques de rehaussement de la parole en matière de PESQ en présence du bruit coloré: rose, bleu, rouge et violet. Les figures, IV.19, IV.20, IV.21 et IV.22 montrent la comparaison des différentes approches en matière de PESQ pour les bruits roses, violets, bleus et rouges respectivement. À partir de ces figures, on peut conclure que l'approche d'Erkelens et al. (2008) [128] donne le meilleur score de PESQ par rapport aux autres techniques.

IV.5.6.4 Comparaison en termes du temps d'exécution

Nous comparons les méthodes mentionnées plus haut en termes du temps d'exécution. Tableau IV.9 montre les résultats des simulations en matière de temps d'exécution, où nous pouvons observer que l'approche géométrique de Yang Lu, Philipos C. Loizou (2008) [131] a moins de temps d'exécution que les autres algorithmes (nous avons utilisé un ordinateur portable Intel (R) Core (TM) i5-3210M CPU @ 2,5 GHz 2.50GHZ).

IV.5.6.5 Application sur notre système de RAL

Nous vérifions l'effet des méthodes de rehaussement de la parole en termes de taux d'identification sur notre système fondé sur MFCCAR et SAD. Dans notre simulation, nous avons utilisé la base de données TIMIT, en définissant les quatre premiers énoncés pour chaque locuteur comme l'ensemble d'apprentissage et 1 énoncé comme l'ensemble de tests. On applique les différentes approches de rehaussement de signal parole (élimination de bruit) sur notre système d'identification et cela en présence d'un bruit WGN en termes de la valeur moyenne du taux d'identification, les résultats de simulation sont rapportés dans le tableau IV.10 dont nous avons choisi 200 locuteurs de la base TIMIT, dont, il est aisément de conclure que l'utilisation de la technique d'Erkelens et al. 2008 [128] donne un meilleur taux d'identification moyenne.

Table IV.1 Valeurs de "α " versus SNR en terme de meilleur taux d'identification.

SNR [dB]	Alpha (α)	Identification rate %
50	0.25	86.00
	0.35	85.66
	0.45	**87.33**
	0.5	85.33
30	0.20	86.33
	0.25	87.00
	0.35	**79.00**
	0.4	77.33
	0.5	76.66
20	0.20	63.66
	0.25	64.33
	0.3	**67.67**
	0.4	57.66
	0.5	42.33
10	**0.25**	**60.33**
	0.3	59.67
	0.35	57.00
	0.4	54.33
	0.5	51.33

Table IV.2 Taux d'identification, le taux d'erreur moyen (HTER) et temps d'exécution moyen en fonction de l'ordre de modèle. (TE Moy=Temps d'exécution moyen).

Ordre GMM / Méthode	32			64			128			256			512		
	FA %	FR %	HTER %	FA %	FR %	HTER %	FA %	FR %	HTER %	FA %	FR %	HTER %	FA %	FR %	HTER %
MFCCAR	5	40	22.5	5	5	5	8	2	5	0	10	5	5	5	5
Id [%]	90			92			96			96			95		
T.E Moy	685.9909														
MFCC	0	80	40	0	85	42.5	0	84	42	0	85	42.5	**20**	**2**	11
Id [%]	73			85			85			88			89		
TE Moy	384.8797														
D–MFCC	0	75	37.5	5	75	40	20	60	40	15	15	15	24	0	12
Id [%]	85				87		90			90			89		
TE Moy	518.3525														
PLP	10	0	5	0	5	2.5	18	1	9	3	12	7.5	6	4	5
Id [%]	84			85			84			86			80		
TE Moy	322.1164														

Tableau IV.3 Taux d'identification par: MFCCAR, MFCC, ΔMFCC et PLP en présence de différentes natures de bruit: WGN, rose, bleu et violet (sans le canal AWGN).

Bruit	SNR [dB]	Le taux d'Identification %			
		MFCCAR	MFCC	ΔMFCC	PLP
WGN	clean	99	85	90	80
	30	93	73	80	71
	20	92	70	80	68
	15	65	50	55	52
	10	50	27	35	30
	5	29	5	15	7
	0	10	5	5	5
Rose	30	95	85	90	78
	20	95	75	83	66
	15	70	55	60	50
	10	55	30	41	21
	5	35	5	15	5
	0	15	7	10	7
Blue	30	88	75	80	71
	20	45	30	32	30
	15	10	15	18	10
	10	6	5	5	5
	5	5	5	5	5
	0	5	5	5	5
violet	30	75	75	70	69
	20	40	30	45	26
	15	15	10	15	7
	10	9	5	5	5
	5	5	5	5	5
	0	5	5	5	5

Tableau IV.4 Taux d'identification moyens par: MFCCAR, MFCC, ΔMFCC, PLP en présence de différentes natures de bruit: WGN, rose, bleu et violet.

Bruit	SNR [dB]	Le taux d'identification moyen %			
		MFCCAR	MFCC	ΔMFCC	PLP
WGN	clean 30 20 15 10 5 0	**62.57**	45.00	51.42	44.71
Rose	clean 30 20 15 10 5 0	**66.28**	48.85	5557	43.57
Bleu	clean 30 20 15 10 5 0	**36.85**	31.42	33.71	29.14
violet	clean 30 20 15 10 5 0	**35.42**	30.85	33.75	27.85

Tableau IV.5 Paramètres de simulation de DS-CDMA.

Paramètres de simulation DS-CDMA	
Code utlisé	PN
Modulation	BPSK
Taille de symbole Ts	2 bits
Temps de chip Tc	1
Canal	Rayleigh
Bande passante	156.25MHZ
Codeur de canal	convolutif ½
Decodeur	Viterbi

Tableau IV.6 Paramètres de simulation d'OFDMA.

Paramètres de simulations OFDM	
IFFT	512
Nombre de sous-porteuese N	100
Espacement entre sous -porteuse Δf	1/640µs = 1.5625Mhz
Bande passante B=N* Δf	1.5625*100=156.25MHZ
Modulation	BPSK
Taille de symbole	2 bits
Taille de mot (wordsize)	8 bits
Nombre de symbole par trame	82
Intervalle de garde (ifft_size/4)	128µs
Durré de symbole = (ifft_size + IG)	640µs
Canal	Rayleigh
Fréquence centrale f_c	128 Hz
Codeur de canal	convolutif ½
Decodeur	Viterbi

Tableau IV.7 BER et Identification du locuteur par: OFDMA et DS-CDMA.

SNR	Techniques d'accès			
	OFDMA		DS-CDMA	
	Id [%]	BER $*10^{-3}$	Id [%]	BER $*10^{-3}$
30	85	0.12012	92	0.00012
25	85	0.20265	92	0.00011
20	80	1.00285	90	0.005824
15	74	1.99520	86	0.12546
10	62	12.20584	76	9.68704
5	15	520.10185	54	448.91701
0	10	612.73529	28	480.49654
-5	5	650.51684	5	501.94012

Tableau IV.8 Mesures moyennes de PESQ pour les méthodes mentionnées précédemment pour la parole contaminée par les bruits de : bavardage, aéroport, voiture, rue et restaurant.

Méthode	SNR [dB]	Types de bruit				
		bavardage	aéroport	voiture	rue	restaurant
Erkelens, et al..2008 [128] (Tracking Of Non-Stationary Noise)	0	**2.3120**	**2.4704**	**2.3750**	**2.3979**	**2.2978**
	5					
	10					
	15					
Yang Lu,et al.. (2008) [131] (Geometric approach to spectral Geometric (GA))	0	2.1820	2.2239	2.0707	2.1328	2.1908
	5					
	10					
	15					
Scalart et al.. (1996) [129] (Speech Enhancement Based On A Priori Signal to Noise estimation)	0	1.7774	1.8102	1.7617	1.9542	1.7062
	5					
	10					
	15					
Cyril Plapous et al.. (2006) [132] (Harmonic Regeneration Noise Reduction - HRNR)	0	1.7562	1.9336	2.0241	1.8911	1.7101
	5					
	10					
	15					
Anthony et al.,(2008) [133]. (Phase Spectrum Compensation - PSC)	0	1.1814	2.2256	2.0742	2.1366	2.1879
	5					
	10					
	15					
Gerkmann et al..(2012) [135]. (Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay)	0	1.7025	2.4206	2.2913	2.2700	2.2839
	5					
	10					
	15					
I. Cohen, 2004 [134].(Speech Enhancement Using a Non causal A Priori SNR Estimator)	0	2.0674	2.0812	2.1271	2.1164	2.0097
	5					
	10					
	15					

Tableau IV.9 Résultats de simulation en termes de temps d'exécution.

Technique de rehaussement de la parole	Temps d'execution [sec]
Erkelens, et al..2008 [128] (Tracking Of Non-Stationary Noise)	0.1972
Yang Lu,et al.. (2008) [131] (Geometric approach to spectral Geometric (GA))	**0.0944**
Anthony et al.,(2008) [133]. (Phase Spectrum Compensation - PSC)	0.0981
Scalart et al.. (1996) [129] (Speech Enhancement Based On A Priori Signal to Noise estimation)	0.7808
Gerkmann et al..(2012) [135]. (Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay)	0.2058
Cyril Plapous et al.. (2006) [132] (Harmonic Regeneration Noise Reduction-HRNR)	0.6113
I.Cohen, 2004 [134]. (Speech Enhancement Using a Non causal A Priori SNR Estimator)	1.3023

Tableau IV.10 Comparaison des taux d'identification moyens en utilisant les différentes méthodes de rehaussement de signal parole.

Method	SNR	Identification Rate [%]	Identification Average [%]
Erkelens, et al..2008 [128] (Tracking Of Non Stationary Noise)	-5	1	51.55
	0	7	
	5	13	
	10	30	
	15	53	
	20	77	
	25	87	
	30	97	
	40	99	
Yang Lu,et al..(2008) [131] (Geometric approach to spectral Geometric GA))	-5	2	45.33
	0	7	
	5	20	
	10	30	
	15	47	
	20	70	
	25	73	
	30	77	
	40	82	

Scalart et al.. (1996) [129] (Speech Enhancement Based on A Priori Signal to Noise estimation)	-5	1	37.55
	0	2	
	5	6	
	10	23	
	15	33	
	20	50	
	25	63	
	30	73	
	40	87	
Cyril Plapous et al.. (2006) [132] (Harmonic Regeneration Noise Reduction-HRNR)	-5	3	38.55
	0	7	
	5	3	
	10	17	
	15	37	
	20	43	
	25	70	
	30	77	
	40	90	
Anthony et al.,(2008) [133]. (Phase Spectrum Compensation - PSC)	-5	2	35.11
	0	5	
	5	7	
	10	10	
	15	13	
	20	43	
	25	60	
	30	77	
	40	99	
Gerkmann et al..(2012) [135]. (Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay)	-5	6	50.77
	0	7	
	5	13	
	10	37	
	15	53	
	20	63	
	25	83	
	30	97	
	40	98	
I.Cohen, 2004 [134]. (Speech Enhancement Using a Non causal A Priori SNR Estimator)	-5	7	36.44
	0	10	
	5	16	
	10	23	
	15	43	
	20	50	
	25	53	
	30	63	
	40	63	

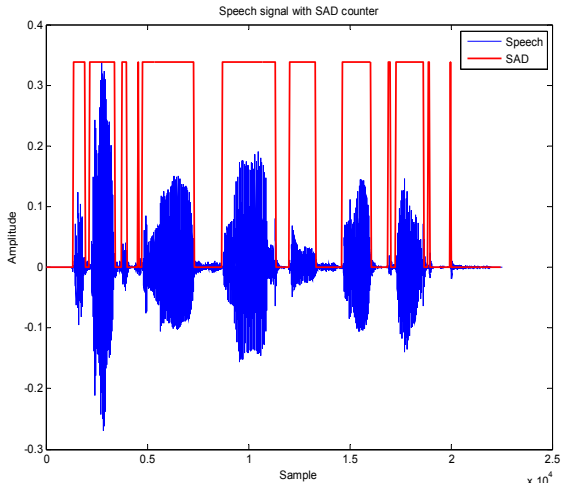


Figure IV.6 Signal parole de la base de données NOIZEUS sans bruit et son contour d'activité de la parole

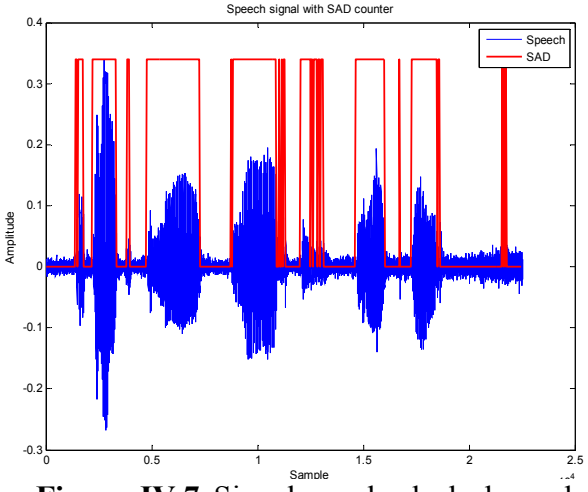


Figure IV.7 Signal parole de la base de données NOIZEUS émergé dans un bruit de bavardage et son contour d'SAD pour SNR =15.

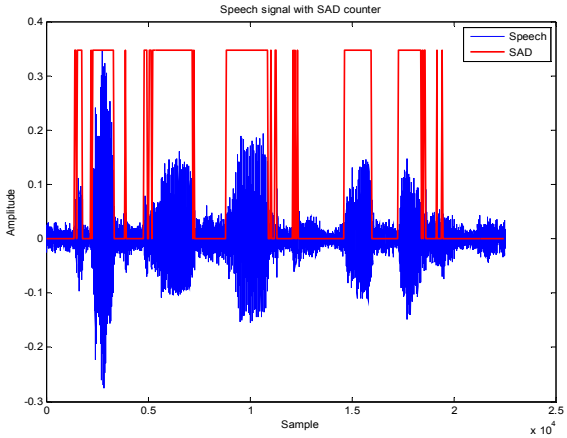


Figure IV.8 Signal parole de la base de données NOIZEUS émergé dans un bruit de bavardage et son contour d'SAD pour SNR =10.

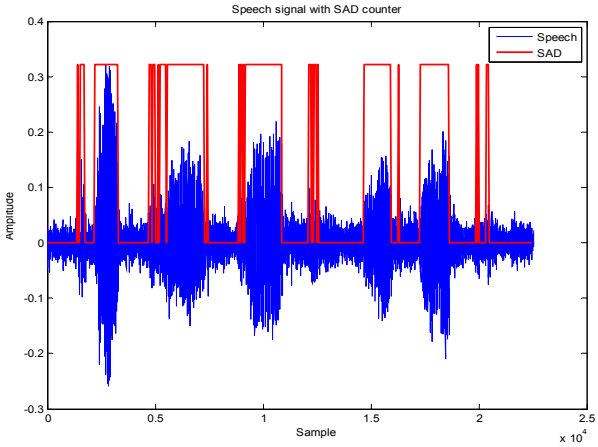


Figure IV.9 Signal parole de la base de données NOIZEUS émergé dans un bruit de bavardage et son contour d'SAD pour SNR =5.

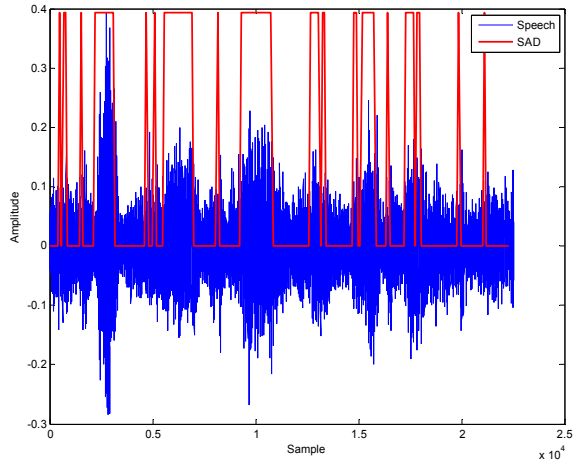


Figure IV.10 Signal parole de la base de données NOIZEUS émergé dans un bruit de bavardage et son contour d'SAD pour SNR = 0.

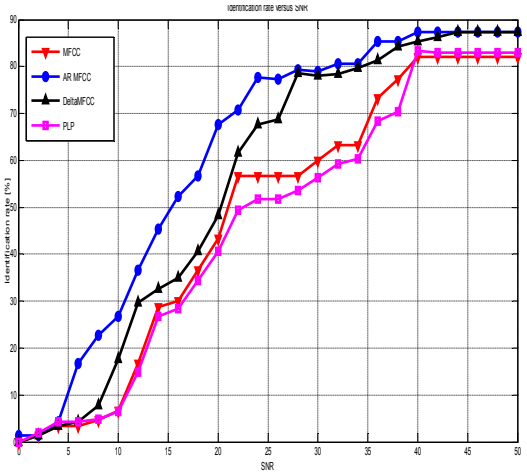


Figure IV.11 Taux d'identification d'ARMFCC, MFCC, PLP et ΔMFCC versus SNR à travers le canal AWGN.

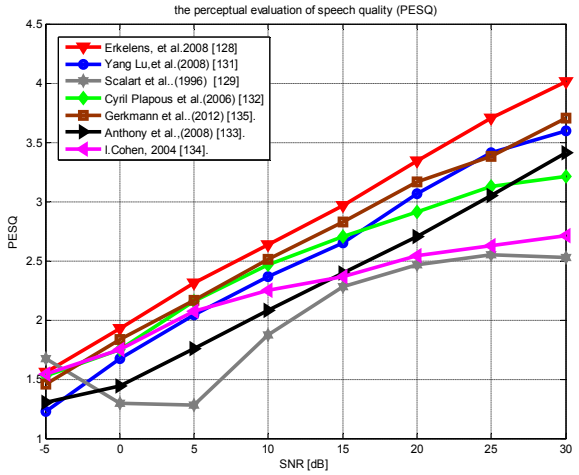


Figure IV.12 Comparaison des performances en termes de PESQ en présence du bruit blanc.(SNR=-5 à 30 dB par pas de 5 dB

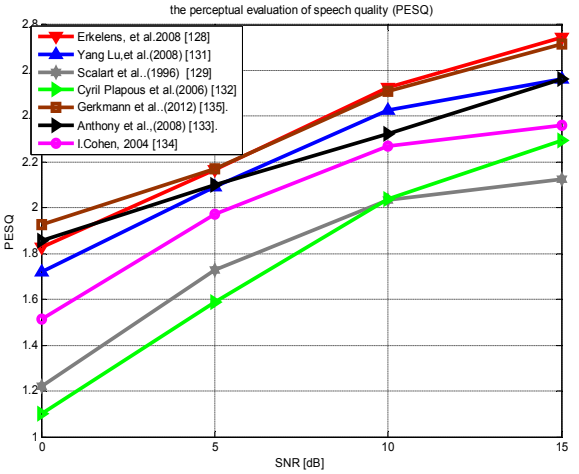


Figure IV.13 Comparaison des performances en termes de PESQ en présence du bruit de bavardage. (SNR=0 dB à 15 dB par pas de 5 dB).

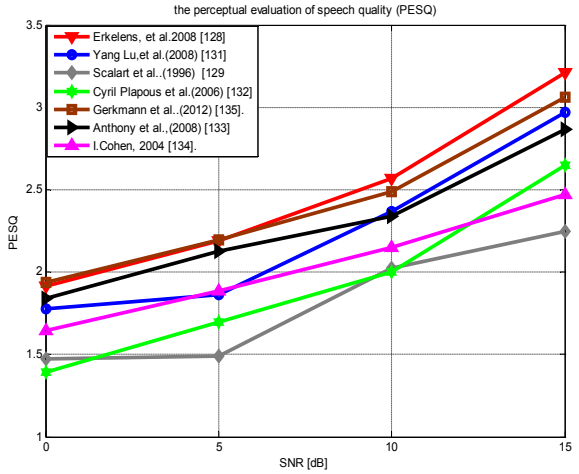


Figure IV.14 Comparaison des performances en termes de PESQ en présence du bruit d'aéroport. (SNR=0 dB à 15 dB par pas de 5 dB).

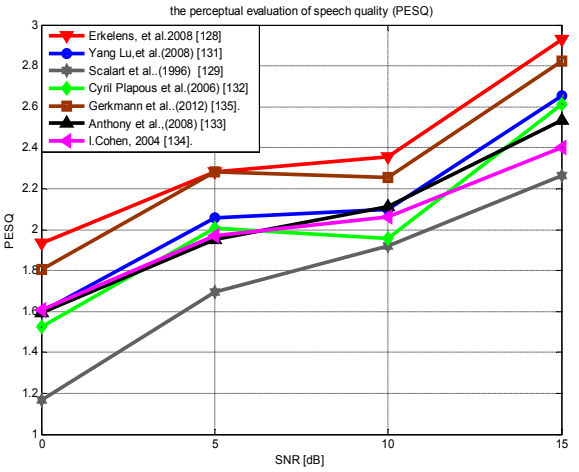


Figure IV.15 Comparaison des performances en termes de PESQ en présence du bruit de voiture. (SNR= 0 dB à 15 dB par pas de 5 dB).

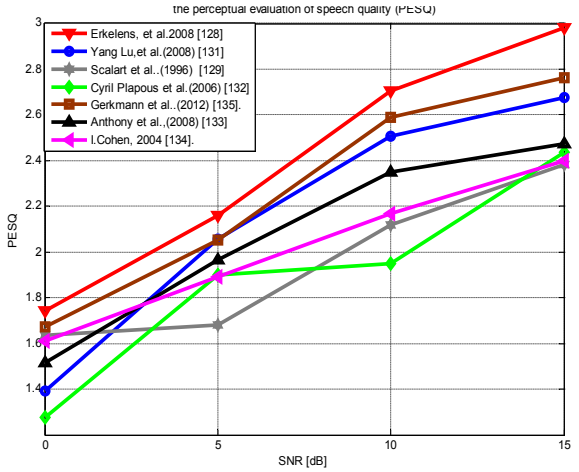


Figure IV.16 Comparaison des performances en termes de PESQ en présence du bruit de la rue. (SNR= 0 dB à 15 dB par pas de 5 dB).

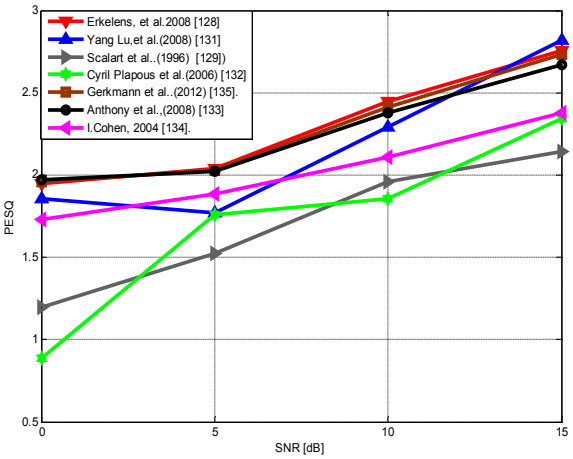


Figure IV.17 Comparaison des performances en termes de PESQ en présence du bruit du restaurant. (SNR= 0 dB à 15 dB par pas de 5 dB).

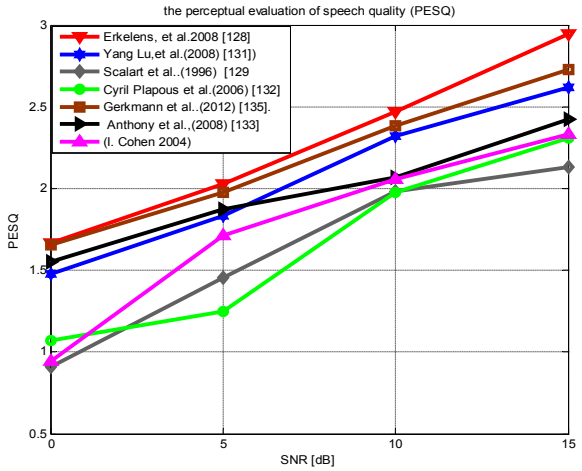


Figure IV.18 Comparaison des performances en termes de PESQ en présence du bruit de salle d'exposition.

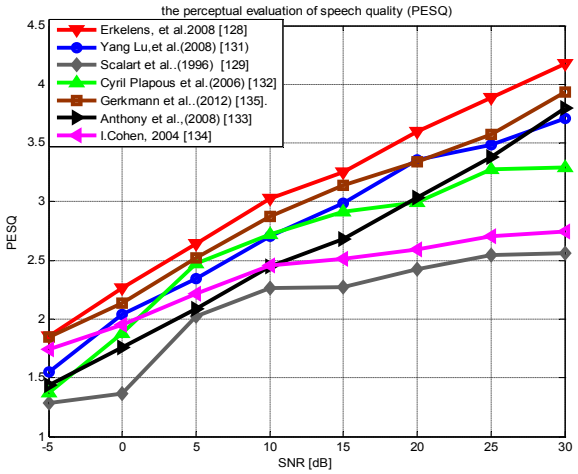


Figure IV.19 Comparaison des performances en termes de PESQ en présence du bruit Rose.

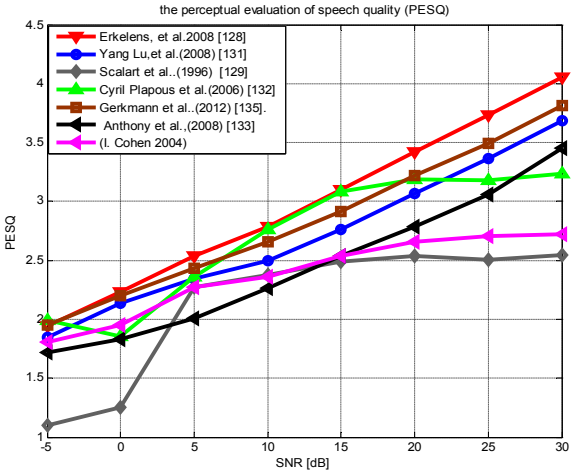


Figure IV.20 Comparaison des performances en termes de PESQ en présence du bruit Violet.

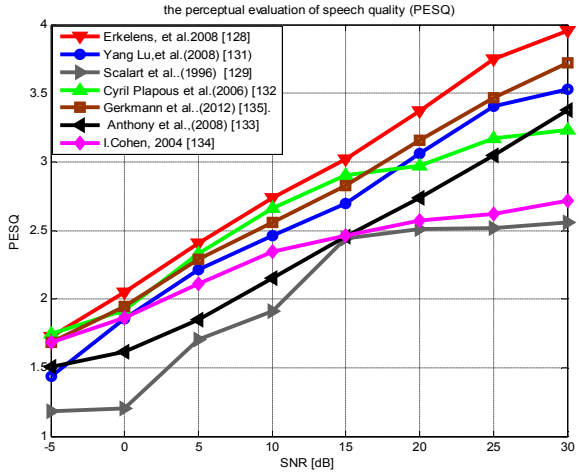


Figure IV.21 Comparaison des performances en termes de PESQ en présence du bruit Bleu.

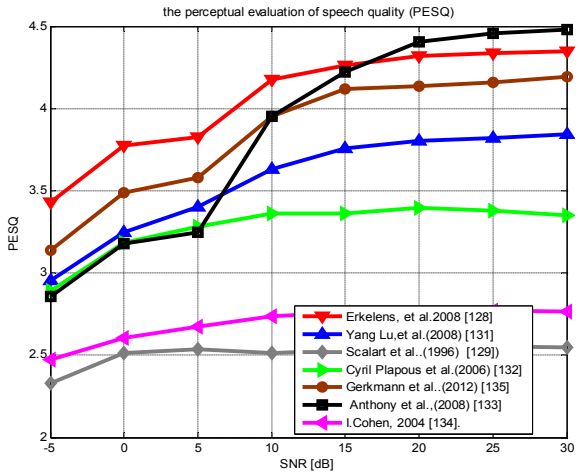


Figure IV.22 Comparaison des performances en termes de PESQ en présence du bruit Rouge.

IV.6 Conclusion

Pour améliorer les performances d'un système de reconnaissance automatique du locuteur à travers le canal AWGN dans un environnement bruité, dans ce chapitre, nous avons fourni une nouvelle approche d'extraction des paramètres robustes fondée sur les coefficients MFCC et les paramètres autorégressifs (AR). Le SAD proposé dans le chapitre précédent dépend d'un facteur « α », dont « α » dépend de niveau de bruit, nous devrions augmenter la valeur de α avec l'augmentation du niveau SNR.

Nous avons proposé une amélioration de l'algorithme SAD vu dans le chapitre précédent par une estimation a priori du niveau de bruit avant la décision parole/silence. Le SAD est basé sur un seuil adaptatif pour la détermination parole/non-parole. Les résultats de simulations ont montré une haute performance de discrimination parole/non-parole dans un environnement bruité justifié par un contour approprié de l'activité de la parole, ce qui améliore le taux de reconnaissance et permet de réduire significativement la dimension des vecteurs acoustiques, le temps de calcul et la mémoire nécessaire pour le stockage des paramètres de la phase d'apprentissage. En outre, il fonctionne correctement dans des environnements d'SNR basse et conduit à un taux de reconnaissance élevé.

L'augmentation du nombre de gaussiennes dans la représentation du locuteur apporte une amélioration du taux d'identification, mais le temps de calcul augmente considérablement. Il est intéressant de noter que le choix de l'ordre du modèle dépend de la quantité de données (la longueur du signal parole) d'apprentissage. Choisir un ordre trop ou peu élevé va nuire la précision du modèle. Pour un nombre de modèles petits, les performances d'identification sont mauvaises dans le cas des coefficients MFCC, mais suffisamment bien pour les coefficients MFCCAR, ΔMFCC (dérivée première de MFCC) et PLP. Le temps d'exécution est moins petit pour le cas de PLP mais plus grand pour MFCCAR.

Le taux d'erreur moyen (HTER) est comprise entre 5% et 42% pour le cas des systèmes de la reconnaissance fondés sur MFCCAR, MFCC, ΔMFCC et PLP en, utilisant un seuil de vérification fixe : $\theta = -2,5129.10^3$. Les taux d'identifications sont meilleurs pour le cas de MFCCAR que MFCC, ΔMFCC et PLP.

Une étude comparative de MFCC (coefficients de 32), ΔMFCC (32), PLP et MFCCAR (64) a été fait compte tenu de leurs effets sur le taux d'identification à distance. Les résultats des expériences ont indiqué que le taux d'identification du locuteur est amélioré en combinant les paramètres AR et MFCC, cependant, en termes du temps d'exécution, MFCCAR nécessite plus de temps que MFCC, PLP et ΔMFCC. Nous avons évalué notre

système de reconnaissance (basé sur MFCCAR et notre SAD) en présence de différents types de bruit comme bruit rose, bleu et violet, mais sans le canal de communication. Les résultats montrent qu'un maximum du taux d'identification (99 %) a été observé pour MFCCAR.

L'utilisation de DS-CDMA donne un taux d'identification meilleur que le système OFDMA et cela à cause que le BER (bit error rate) avec le DS-CDMA est mois que celle d'ODFMA. Ainsi la remarque la plus importante est que le système de reconnaissance base sur DS-CDMA est plus robuste au bruit contrairement au OFDMA et cela justifié par les codes utilisés par le DS-CDMA qui rend le système plus robuste au bruit.

En outre, on a fait une comparative de différentes méthodes de rehaussement de la parole (élimination de bruit) en présence de plusieurs types de bruit (bruits colorés et bruits réels). L'évaluation a été effectuée par l'évaluation perceptive des scores de la qualité de la parole PESQ (l'UIT-T P.862). L'approche de Erkelens et al. (2008) [128] a donné un PESQ significative différents niveaux d'SNR. En termes du temps d'exécution, l'approche géométrique proposée par Yang Lu,et al..(2008) [131] nécessite moins de temps d'exécution que les autres approches. D'autre part, l'application des approches de réhaussement de la parole sur notre système d'identification basée sur MFCCAR et SAD a donné des taux d'identification importante en utilisant la technique d'Erkelens et al. 2008 [128].

Notre système d'identification du locuteur peut être très efficace en diminuant le temps d'exécution de MFCCAR et développant une technique de rehaussement de signal parole.

Conclusion générale et perspectives

Une des caractéristiques plus fascinantes de l'homme est leur capacité de communiquer des idées au moyen de la parole. Cette capacité est sans aucun doute l'un des faits qui a permis le développement de notre société. L'homme a été toujours attiré par la possibilité de créer des machines capables de produire le signal parole et de reconnaissance du locuteur.

Un système de reconnaissance automatique du locuteur (RAL) peut être défini comme un mécanisme capable de décoder le signal produit dans les voies nasales et vocaux d'un locuteur humain dans la séquence d'unités linguistiques contenues dans le message que le locuteur désire communiquer.

L'objectif final de la RAL est la communication homme-machine. Cette façon naturelle de l'interaction a trouvé de nombreuses applications en raison du développement rapide de différentes techniques matérielles et logiciel. Les plus importants sont l'accès aux systèmes d'information, aide aux handicapés et le contrôle du système par le signal parole.

La présente thèse est axée sur une large classe d'application qui implique l'accès par la reconnaissance du locuteur à des systèmes ou services d'information à distance (RSR). Ces types d'applications ont été clairement soutenus par le développement rapide des réseaux numériques (cellulaires et Internet) lors des 15 dernières années. Ces systèmes comportent une architecture client-serveur dans laquelle le serveur contient les informations et le client choisit la parole, qui est transmise au serveur sous une forme appropriée.

Le sujet de RSR implique une connaissance a priori de traitement de la parole, (caractéristique spectrale, codage, technique d'extraction de paramètres et le prétraitement de signal parole…) et télécommunications (réseau internet, mobile et réseaux de communication). Bien que, le signal de parole soit un processus aléatoire non stationnaire à long terme dans le chapitre I nous avons décri les différentes classes des paramètres de l'analyse acoustique. Un système de reconnaissance automatique du locuteur, quelle que soit la tâche considérée, se résume en trois étapes principales : l'analyse acoustique du signal de

parole, la modélisation du locuteur et la décision. Également, tout système de RAL dépend de la technique d'extraction de paramètres utilisé, modélisation, décision, et ainsi la phase de prétraitement..

Le sujet de cette thèse nous a exigé à présenter une vue d'ensemble sur la reconnaissance du locuteur sur les réseaux mobile, sans fil et internet. On a donné un aperçu sur le signal parole dans les réseaux (Mobile et IP). La première conclusion est qu'il y ait deux architectures qu'on peut utiliser pour la mise en œuvre d'un système de RSR sur un canal numérique: dans la première approche, généralement connue comme la reconnaissance du locuteur/speech dans les réseaux (Network Speaker/speech Recognition NSR), le système de reconnaissance réside dans le réseau de la perspective du client. Dans ce cas, la parole est compressée par un codec (codeur-décodeur) de la parole afin de permettre une transmission de faible débit binaire et/ou d'utiliser un canal du trafic existant des paroles (comme dans le cas de la téléphonie mobile). Bien qu'il soit également possible d'extraire les caractéristiques de reconnaissance directement à partir des paramètres du codec, c'est la deuxième approche connue sous le nom de reconnaissance du locuteur/speech distribué (distributed speech/speaker recognition-DSR). Dans notre travail (Thèse), on a adopté la RAL conformément à NSR.

Nous avons proposé une architecture de notre système de reconnaissance à distance (RSR) qui se base sur une nouvelle approche de détection parole/non-parole (speech activity détection-SAD) ce qui signifie une amélioration de la capacité mémoire dont l'extraction des paramètres ne s'exerce pas dans les zones classifiées comme des zones de silence qui vont conduire à un taux d'identification amélioré. Notre algorithme proposé SAD qui basait sur l'énergie et le taux de passage par zéro, donne un contour approprié de l'activité de la parole. En outre, il a fonctionné avec précision dans des environnements de faible SNR (jusqu'à SNR de 5 dB) et conduit à une bonne amélioration du taux d'identification.

Bien que le type de codage ou de compression fût un facteur important, nous avons fait une étude comparative des codecs vocaux: PCM, DPCM et ADPCM en, tenant compte de leurs effets sur la performance de notre système de reconnaissance automatique distant et cela dans un environnement bruité. La meilleure performance globale de codecs vocaux a été observée pour le code de PCM en termes du taux de reconnaissance et de temps d'exécution. Il est

recommandé alors d'utiliser la technique PCM comme codec de la parole dans les systèmes de reconnaissance du locuteur à distance destiné pour les applications VoIP.

Les performances d'un système de reconnaissance automatique du locuteur à distance se dégradent à cause de distorsion du canal, présence de bruits et le codage ou la compression de signal parole destiné pour la transmission. Afin d'améliorer le taux de reconnaissance, l'utilisation d'un codage de canal est nécessaire pour rendre le système à distance plus robuste contre les erreurs de canal; par conséquent, nous avons choisi de code convolutif.

Dans ce travail, on a développé une nouvelle technique d'extraction de caractéristiques de la parole fondée sur les coefficients MFCC et les paramètres autorégressifs (AR) ce qui nous a donné une amélioration des performances de notre système de reconnaissance automatique du locuteur à travers le canal AWGN dans un environnement bruité. La connaissance a priori de niveau de bruit (SNR) est nécessaire pour une décision performante de parole/non-parole.

Nous avons proposé un algorithme de détection d'activité vocale (SAD) efficace basé sur une connaissance du niveau de bruit (SNR). Le SAD est basé sur un seuil pour la détermination du parole/non-parole. L'algorithme SAD a montré une haute performance de discrimination de la parole/non-parole dans un environnement bruité ce qui améliore le taux de reconnaissance et permet de réduire significativement la dimension des vecteurs acoustiques et le temps de calcul et la mémoire nécessaire pour le stockage des paramètres de la phase d'apprentissage. Le SAD dépend d'un facteur « α », où pour avoir un taux élevé d'identification, nous devrions augmenter α avec l'augmentation du niveau SNR. SAD a donné un contour approprié de l'activité de la parole.

L'augmentation du nombre de gaussiennes dans la représentation du locuteur apporte une amélioration du taux d'identification, mais le temps de calcul augmente considérablement. Mais il est intéressant de noter que le choix de l'ordre du modèle dépend de la quantité de données (la longueur du signal parole) d'apprentissage. Choisir un ordre trop peu élevé va nuire la précision du modèle. Ainsi choisir, un trop de composantes engendreront une charge de calcul plus importante.

Une comparaison des techniques d'extraction des paramètres PLP, MFCC, ΔMFCC et notre approche MFCCAR est faite. Le temps d'exécution est moins petit pour le cas de PLP mais plus grand pour MFCCAR. En utilisant un seuil de vérification fixe $\theta = -2.5129.10^{3}$, le

taux d'erreur moyen (HTER) est comprise entre 5% et 42%. Les taux d'identification sont meilleurs pour le cas de notre RSR basé sur MFCCAR que les systèmes basés sur MFCC, ΔMFCC et PLP. En d'autres termes pour un nombre de modèles petits (modèle GMM), les taux d'identifications sont mauvais pour les coefficients MFCC, mais suffisamment bien pour le cas des coefficients MFCCAR, ΔMFCC et PLP.

Les résultats des expériences ont indiqué que le taux d'identification est amélioré en combinant les paramètres AR et MFCC. Cependant, en termes du temps d'exécution, le MFCCAR nécessite plus de temps que: MFCC, ΔMFCC et PLP.

L'application des techniques d'accès multiple OFDM et DS-CDMA sur notre système de RSR sur le canal Rayleigh a donné un meilleur taux d'identification pour le DS-CDMA que le système OFDMA et cela justifié par le faible BER pour DS-CDMA par rapport au celle d'ODFMA. Ainsi, la remarque la plus importante est que le système de reconnaissance avec le DS-CDMA est plus robuste au bruit contrairement au OFDMA et cela justifié par les codes (séquence de codes "DS") utilisés par le DS-CDMA qui rend le système plus robuste au bruit.

Aussi, nous avons fait une étude comparative de sept méthodes d'amélioration de la parole, compte tenu de leurs effets sur le taux de reconnaissance de notre système de reconnaissance du locuteur. La comparative a montré que la méthode proposée par Erkelens et al. (2008) a fourni la bonne précision de taux d'identification, de sorte que nous recommandons cette méthode pour améliorer signal de parole.

Notre système d'identification du locuteur peut être très efficace en diminuant le temps d'exécution d'MFCCAR et le développement d'une nouvelle technique de rehaussement de signal parole.

Comme principales perspectives dans ces domaines de recherche: la première observation que nous pouvons faire sont que toutes les contributions, présentées dans ce travail, ont été évaluées dans un cadre simulé, une phase de validation en conditions réelles de fonctionnement est encore nécessaire. La recherche se continue dans la recherche des autres techniques d'extraction des paramètres. Ainsi, l'amélioration du taux de reconnaissance par le développement d'autres techniques de rehaussement adaptés à notre application.

Annexes

Annexe A: Listes des contributions scientifiques

Publications scientifiques internationales

1. Riadh AJGOU, Salim SBAA, Said GHENDIR, Ali CHEMSA, A. Taleb-Ahmed,″ Robust Speaker Identification System Over AWGN Channel Using Improved features Extraction and Efficient SAD Algorithm with Prior SNR Estimation″, *international journal of circuits, systems and signal processing*, 2016, vol. 10.pp. 108-118.
2. Riadh AJGOU, Salim SBAA, Said GHENDIR, Ali CHEMSA, A. TALEB-AHMED,″ An Efficient Approach for MFCC Feature Extraction for Text Independant Speaker Identification System″, *international journal of communiucations*, 2015, vol. 09.pp.114-122.
3. Riadh AJGOU, Salim SBAA, Said GHENDIR, Ali CHEMSA, A. TALEB-AHMED, ″ Novel Detection Algorithm of Speech Activity and the impact of Speech Codecs on Remote Speaker Recognition System″, *WSEAS Transactions on Signal Processing*, 2014, vol. 10. pp. 309-319.

Publications scientifiques nationales

1. AJGOU R., SBAA S., AOURAGH S, et TALEB-AHMED A. Détection du pitch par les ondelettes continues en temps réel pour un signal parole basé sur un seuil adaptatif pour unes détermination V/NV. Courrier du savoir, 2012, N°12, Octobre 2011, pp.21-26.

Communications internationales

1. Ajgou Riadh, Sbaa Salim, Said. Ghendir, Chemsa Ali, Abdelmalik Taleb-Ahmed, " New Speech Enhancement Method based on Wavelet Transform and Tracking of

Non Stationary Noise Algorithm," *Proceedings of Recent Advances on Electroscience and Computers, Barcelona,* Spain, April 7-9, 2015.pp 45-52.

2. Riadh Ajgou, Salim Sbaa, Said Ghendir,Ali Chamsa and A. Taleb-Ahmed," Speaker Recognition System Based on ARMFCC and SAD Algorithm with Prior SNR Estimation and Adaptive Threshold over AWGN channel," *Proceedings of the International Conference on Recent Advances in Electrical Engineering and Educational Technologies ,* Athens, Greece, November 28-30, 2014. pp 120-128.
3. AJGOU, Riadh, SBAA, Salim, GHENDIR, Said, and A. Taleb-ahmed.. Robust remote speaker recognition system based on AR-MFCC features and efficient speech activity detection algorithm. In : *Wireless Communications Systems (ISWCS), 2014 11th International Symposium on*. IEEE, 2014. p. 722-727
4. Riadh Ajgou, Salim Sbaa, Said Ghendir,Ali Chamsa and A. Taleb-ahmed," Effects of speech codecs on a remote speaker recognition system using a new SAD," *Proceedings of the 2014 International Conference on Systems, Control, Signal Processing and Informatics II (SCSI '14),* Prague, Czech Republic April 2-4, 2014. pp 71-78.
5. R. Ajgou, S. Sbaa, S. Ghendir, A. Taleb-Ahmed, "Détection du pitch par les ondelettes discrètes en temps réel par un seuil adaptatif ", *Third International Conférence on Image ans Signal Processing and their Applications*, Mostaganem university, ISPA 2012.
6. R. Ajgou, S. Sbaa, S. Ghendir, A. Taleb-Ahmed , "Détection du pitch par un seuil adaptatif et en temps réel par les ondelette discrètes, " *First International Conference on Signal, Image, Vision and their applications,* Guelma, Algeria, Novembre 21-24, 2011. P.187-192
7. AJGOU, R., SBAA, S., AOURAGH, S., & A. TALEB-AHMED "Détection de pitch en temps réel basée sur les Ondelettes continue pour un signal parole basée sur un seuil adaptatif pour une détermination V/NV". Second International Conférence on Image ans Signal Processing and their Applications, Biskra university, ISPA 2010.

Annexe B : OFDM

B.1 Introduction :

Un des problèmes majeurs en télécommunications est d'adapter l'information à transmettre au canal de propagation. Pour des canaux sélectifs en fréquence, une technique est l'utilisation de modulations multi-porteuses dans laquelle un bloc d'information est modulé par une transformée de Fourier. Cette technique connue sous le nom d'OFDM a connu un vif succès ces dernières années et est en phase de normalisation dans différents standards des réseaux sans fils (IEEE802.11a, WiMAX, LTE, …). La technique OFDM a le grand mérite de transformer un canal multi-trajet large bande en un ensemble de sous-canaux mono-trajet très simples à égaliser. De plus, l'utilisation ingénieuse de redondance cyclique à l'émission permet de réduire la complexité des terminaux grâce à l'utilisation d'algorithmes à base de FFT rapides [124].

B.2 Modulations Multi-porteuses

Dans le cas d'un canal à trajets multiples, les techniques de modulation classiques sont très sensibles à l'interférence inter-symboles (intersymbol interference ou ISI). Cette interférence est d'autant plus importante que la durée d'un symbole est petite par rapport au delay spread du canal. L'intérêt des modulations multi-porteuses (Multi-Carrier Modulation) est de placer l'information dans une fenêtre temps-fréquence telle que sa durée soit bien plus grande que le delay spread du canal de propagation. Cette avantage, primordial pour les communications sans fils, en fait une solution pressentie pour les différents types de réseaux haut débit sans fils: réseaux cellulaires, réseaux locaux sans fils et boucle locale radio. L'idée originale des modulations multi-porteuses est de transformer l'étape d'égalisation dans le domaine temporel par une égalisation simplifiée dans le domaine fréquentielle pour retrouver le signal émis. Afin de décrire le principe, considérons un circuit électrique pour lequel la réponse du courant (ici, le signal émis) est régie par une équation différentielle [124].

B.3 Principe

Le principe de l'OFDM consiste à répartir sur un grand nombre de sous-porteuses le signal numérique que l'on veut transmettre. Comme si l'on combinait le signal à transmettre sur un grand nombre de systèmes de transmission (des émetteurs, par exemple) indépendants et à des fréquences différentes.

Pour que les fréquences des sous-porteuses soient les plus proches possibles et ainsi transmettre le maximum d'information sur une portion de fréquences donnée, l'OFDM utilise des sous-porteuses orthogonales entre elles. Les signaux des différentes sous-porteuses se chevauchent mais grâce à l'orthogonalité n'interfèrent pas entre eux.

En codage orthogonal, l'espacement entre chaque sous-porteuse doit être égal à [125]:

$$\Delta f = \frac{k}{Tu}[hz] \tag{B.1}$$

Où T_U [sec] est la durée utile d'un symbole (c.à.d. la taille de la fenêtre de capture du récepteur), et k est un entier positif, généralement égal à 1. (Un exemple simple: Une durée de symbole utile TU = 1 ms exigerait un espacement de sous-porteuse de $\Delta f = \frac{1}{1ms} = 1KHZ$) Par conséquent, avec N sous-porteuses, la largeur totale de la bande passante sera de [125]:

$$B \approx N.\Delta f[hz] \tag{B.2}$$

L'orthogonalité permet également une haute efficacité spectrale . Le multiplexage orthogonal produit un spectre de fréquence presque plat (typique du bruit blanc), ce qui entraîne un minimum d'interférences avec les canaux adjacents. U-n filtrage séparé de chaque sous-porteuse n'est pas nécessaire pour le décodage, une transformée de Fourier FFT étant suffisante pour séparer les porteuses entre elles.

Le signal à transmettre est généralement répété sur différentes sous-porteuses. Ainsi dans un canal de transmission avec des chemins multiples où certaines fréquences seront détrités à cause de la combinaison destructive de chemins, le système sera tout de même capable de récupérer l'information perdue sur d'autres fréquences porteuses qui n'auront

pas été détruites. Chaque sous-porteuse est modulée indépendamment en utilisant des modulations numériques : BPSK, QPSK, QAM-16, QAM-64,…

Ce principe permet de limiter l'interférence entre symboles. Pour l'éliminer, on peut ajouter un intervalle de garde (c'est-à-dire une période pendant laquelle il n'y a aucune transmission) après chaque symbole émis, très grand devant le délai de transmission (la distance séparant l'émetteur du récepteur divisée par la vitesse de la lumière).

Le décodage OFDM nécessite une synchronisation très précise de la fréquence du récepteur avec celle de l'émetteur. Toute déviation en fréquence entraîne la perte de l'orthogonalité des sous-porteuses et crée par conséquent des interférences entre celles-ci. Cette synchronisation devient difficile à réaliser dès lors que le récepteur est en mouvement, en particulier en cas de variation de vitesse, de direction ou si de nombreux échos parasites sont présents.

B.4 Descriptions mathématique

L'équivalent passe-bas d'un signal OFDM est exprimé ainsi [125]:

$$v(t) = \sum_{k=0}^{N-1} I_k e^{i2\pi k\ t/T}, 0 \le t < T \qquad \text{(B.3)}$$

Où I_k sont les symboles de donnée, N est le nombre de sous-porteuses et T la durée du bloc OFDM. L'espacement entre porteuses de $1/T$ [Hz] rend les sous-porteuses orthogonales entre elles ; cette propriété est exprimée ainsi [125,126] :

$$\frac{1}{T}\int_0^T \left(e^{i2\pi k_1 t/T}\right)^* \left(e^{i2\pi k_1 t/T}\right) dt = \frac{1}{T}\int_0^T \left(e^{i2\pi (k2-k2) t/T}\right) dt = \begin{cases} 1, & k_1 = k_2 \\ 0, & k_1 \neq k_2 \end{cases} \qquad \text{(B.4)}$$

Où $(.)^*$ correspond à l'opérateur conjugué complexe.

Pour éviter l'interférence inter-symboles dans un environnement de propagation multichemins, un intervalle de garde $-Tg \le t < 0$, où Tg est la période de garde, est inséré avant le bloc OFDM. Pendant cet intervalle, un *préfixe cyclique* (PC) est transmis. Ce préfixe cyclique est égal au dernier Tg du bloc OFDM. Le signal OFDM avec le cyclique préfixe est donc [125, 126] :

$$v(t) = \sum_{k=0}^{N-1} I_k e^{i2\pi k \ t/T}, -Tg \le t < T \qquad \text{(B.5)}$$

Le signal passe-bas ci-dessus peut soit être constitué de valeur réelles ou complexes. Pour le signal à valeurs réelle celui-ci est généralement transmis en bande de base et exprimé ainsi [125, 126]:

$$s(t) = \Re\left\{v(t) e^{i2\pi f_c t}\right\} \qquad \text{(B.6)}$$

Le signal en bande de base à valeurs complexes est par contre modulé à une fréquence supérieure f_c. En général, le signal est représenté ainsi [125, 126]:

$$s(t) = \sum_{k=0}^{N-1} |I_k| \cos\left(2\pi\left[fc + {k}/{T}\right]t + \arg\left[I_k\right]\right) \qquad \text{(B7)}$$

Annexe C : DS-CDMA

C.1 Introduction

L'étalement de spectre en séquence directe 'DS' (DS-CDMA) utilise l'étalement de spectre par séquence directe afin de permettre la transmission simultanée de signaux issus de plusieurs utilisateurs à l'intérieur d'une même bande de fréquence, tout en assurant un taux d'interférences inter-utilisateurs assez faible [123]. Habituellement, dans un contexte dit "coopératif", le récepteur connaît la séquence d'étalement utilisée par l'émetteur pour étaler le signal qui lui est destiné, ce qui lui permet d'extraire le signal informatif émis à partir du signal reçu. Ceci [123], sans qu'un autre récepteur (utilisateur) puisse en faire de même, car ne connaissant pas la séquence d'étalement en question. Par contre, dans le contexte "non coopératif" auquel nous nous intéressons ici, le récepteur ne connaît pas les séquences d'étalement de l'émetteur. Dans ce contexte, les propriétés intrinsèques de l'étalement de spectre posent de gros problèmes. Même pour une transmission mono-utilisateur (ou il n'y a qu'un seul signal étalé plus du bruit), le signal étalé est souvent caché en dessous du niveau du bruit, ce qui a pour conséquence de masquer la transmission [113]. De plus, le choix des séquences d'étalement (généralement pseudo aléatoires) fait que le signal lui même a des caractéristiques statistiques ressemblant à du bruit et est donc difficile à détecter [123].

C. 2. Etalement de Spectres

L'étalement de spectre consiste à étendre la bande de fréquence du signal à transmettre; la densité spectrale de puissance du signal utile est diminuée. Ce signal est perçu comme un bruit pour un utilisateur non concerné par la transmission. Cela est réalisé grâce à un codage de l'information à transmettre avec une séquence ***pseudo-aléatoire (Pseudo Noise – code, PN code),*** de longueur N, connue seulement des utilisateurs. Dans les systèmes CDMA, les utilisateurs se partagent toute la bande passante de manière continue. On assigne une signature, ou code, à chaque utilisateur de manière à pouvoir les identifier au

récepteur. L'orthogonalité, ou la quasi-orthogonalité, de ces signatures permet d'isoler chacun des canaux [123]. La figure C.1 représente une description de CDMA.

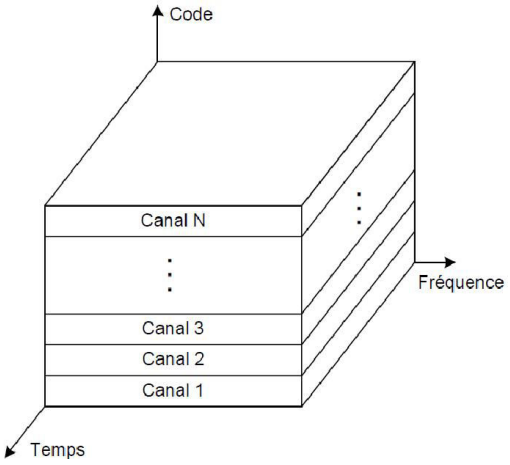


Figure C.1 Description de CDMA, où on associe un code différent à chaque utilisateur [127].

Contrairement au FDMA et au TDMA qui ont un seuil de capacité strict, le CDMA offre un seuil de capacité souple, les performances du système se dégradant graduellement avec l'ajout de nouveaux utilisateurs. Le CDMA permet à chaque utilisateur de profiter de la totalité de la bande passante en tout temps et le codage offre une protection contre les interférents. Cependant, le CDMA est susceptible à l'effet proche-loin lorsque l'utilisateur désiré à une faible puissance relativement à un autre utilisateur. De plus, puisque les codes ne sont généralement pas exactement orthogonaux, les utilisateurs d'un même système s'interfèrent mutuellement.

C.2.1 L'étalement De Spectre En Séquence Directe 'DS'

L'étalement de spectre en séquence directe **(**figure C.2**)** se fait par la multiplication de l'information à transmettre de débit *Rb* par un code pseudo-aléatoire, aussi appelé signature, ayant un débit *Rc*. On a [127]:

$$N = \frac{T_c}{R_b} = \frac{T_b}{T_c} \qquad \text{(C.1)}$$

Où $T_b = \frac{1}{R_b}$ la durée d'un bit d'information $T_c = \frac{1}{R_c}$ est la durée d'une impulsion rectangulaire du code, appelée chip. *N* est habituellement un entier, supérieur à 1 puisqu'il mesure l'étalement du spectre et représente le nombre de chips par bit d'information. On appelle également ce rapport gain de traitement (*processing gain*).

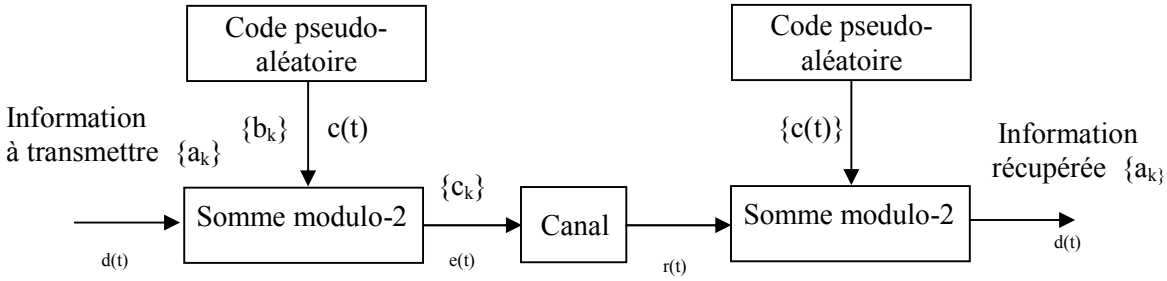


Figure C.2 Schéma général d'un système à étalement de spectre en séquence directe [127].

Le message *d(t)* à transmettre se présent sous forme d'un code binaire non retour à zéro (NRZ) [127] :

$$d(t) = \sum_{k=-\infty}^{\infty} a_k \times h(t - kT) \qquad \text{(C.2)}$$

Les symboles a_k prenant leur valeur dans l'ensemble {-1,1} et h(t) est la fonction porte de durée T. L'étalement de spectre suivant la technique de la séquence directe consiste à multiplier le message m(t) par un signal numérique NRZ c(t) dont le rythme est un multiple de celui du message m(t) [127] :

$$c(t) = \sum_{k=-\infty}^{\infty} b_k \times H(t - kT_c) \qquad \text{(C.2)}$$

Avec $T_c = \frac{T}{N}$ et b_k sont des symboles binaires qui prennent leur valeur dans l'ensemble {-1, 1} et H(t) est la forme du signal (Signal Shape) « fonction porte de durée Tc ». N s'appelle le facteur d'étalement. Après multiplication des signaux d(t) et c(t) on obtient le signal à spectre étalé e(t) [127]:

$$e(t) = d(t).c(t) = \sum_{k=-\infty}^{\infty} c_k \times H(t - kT_c) \qquad \text{(C.2)}$$

Avec $c_k = a_k.b_k$ cette opération d'étalement du spectre est illustrée sur la figure C.3.

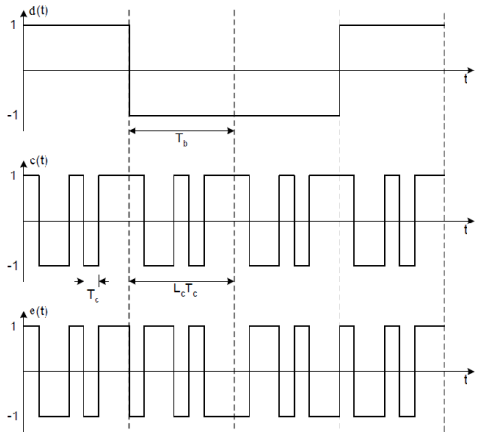


Figure C.3 Principe de l'étalement de spectre (technique de la séquence directe)[127].

La densité spectrale de puissance du signal m(t) et du signal e(t) s'obtient sans difficulté. La Figure C.4 représente la densité spectrale d'un signal à spectre étalé, dont [113] : $\gamma_m(f) = T\left[\frac{\sin \pi fT}{\pi fT}\right]^2$ et $\gamma_c(f) = T_c\left[\frac{\sin \pi fT}{\pi fT}\right]^2$.

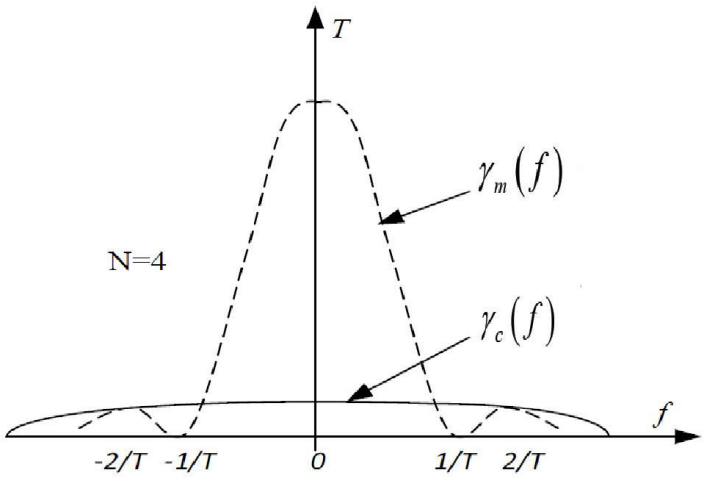


Figure C.4 : densité spectrale d'un signal à spectre étalé [123].

C. 3 Choix de Code d'étalement

Il ya plusieurs techniques qui existent pour construire des codes ayant de bonnes propriétés.

C.3.1 PN séquence code :

Si on veut disposer d'un grand nombre de signatures, suffisamment, orthogonales entre elles, on doit faire en sorte qu'elles présentent beaucoup de transitions, répartitions de manière aléatoire elles doivent donc ressembler à un « bruit binaire ». Cette technique est associée aux générateurs pseudos aléatoires, on parle alors de PN séquences, le préfixe PN signifiant pseudo-noise. Toutes les séquences pseudo-aléatoires **(**Ls séquences à longueur maximale : m-séquences) sont habituellement générées par un *registre à décalage avec rétro-action linéaire* (*linear feedback shift register*) [127].

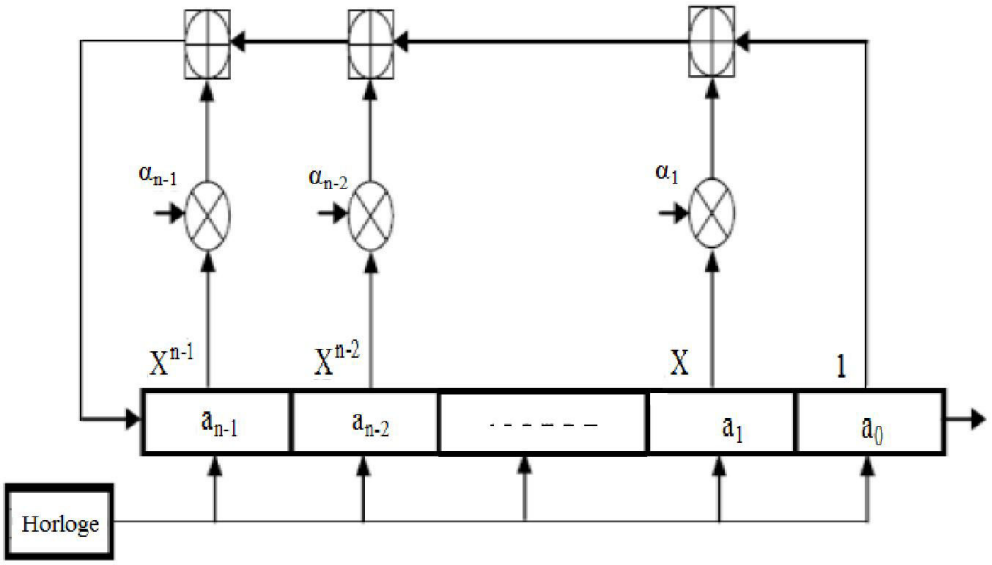


Figure C.5 Schéma générique d'un registre à décalage [123]

Un registre à décalage binaire, comme celui-ci décrit par la figure C.5, représente l'une des manières les plus courantes pour générer des codes pseudo-aléatoires. Son fonctionnement est le suivant, Une fois initialisés les différents états du registre, le bit en sortie est calculé à chaque coup d'horloge en additionnant **modulo 2** tous les bits présents à chaque état. Les bits sont ensuite décalés de manière circulaire pour réinitialiser les états et calculer le bit suivant. Le registre à décalage est dit périodique, car quelles que soient les valeurs initiales, c'est-à-dire les valeurs prises par « *ai* », on retrouve ces mêmes valeurs après un nombre fini de périodes d'horloge. Comme le registre comprend « n »

états représentés par les valeurs binaires de « ai », il est possible de générer « *2n* » codes pseudo-aléatoires. Il en résulte aussi que la période de la séquence n'est jamais supérieure à «*2n*-1 ». En outre, on peut voir sur la figure C.5, que lorsque les valeurs initiales sont toutes égales à zéro, le registre reste dans le même état de façon permanente [123] :

Le nombre maximum d'états possibles, différents de « 0 », est « $m = 2^n - 1$ ».

 Une séquence binaire de période « $m = 2^n - 1$ », générée avec un registre à décalage de type LFSR (Un registre à décalage à rétroaction linéaire [127]), est appelée m-séquence ou encore, séquence à longueur maximale (Maximal Length Sequence [123].

 Les coefficients « αi » peuvent prendre deux valeurs « *1* » ou « *0* ». Lorsqu'il y a une connexion physique, « $\alpha i = 1$ » et lorsque « $\alpha i = 0$ », il n'y a pas de connexion

.

C.3.2 Les codes de Gold :

Les séquences de Gold sont utiles en raison de grand nombre de codes disponibles, leur principale qualité est liée au fait que la fonction d'inter-corrélation entre deux codes est uniforme et bornée, ceci les rends utiles pour les techniques d'accès multiples [123].

Bibliographie

[1] Mami, Yassine. *Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence*. Diss. Télécom ParisTech, 2003.

[2] AL-SAWALMEH, Wael, DAQROUQ, Khaled, AL-QAWASMI, Abdel-Rahman, *et al.* The use of wavelets in speaker feature tracking identification system using neural network. *WSEAS Transactions on Signal Processing*, 2009, vol. 5, no 5, p. 167-177.

[3] Teva MERLIN .Amiral, une plateforme générique pour la reconnaissance automatique du locuteur de l'authentification à l'indexation. (thèse) Académie d'aix-Marseille université d'Avignon et des Pays de Vaucluse. 18 novembre 2004.

[4] TIWARI, Vibha. MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, 2010, vol. 1, no 1, p. 19-22.

[5] IMPEDOVO, Donato et REFICE, Mario. Frame length selection in speaker verification task. *Transaction on Systems*, 2008, vol. 7, no 10, p. 1028-1037.

[6] MEIGNIER, S. Indexation en locuteurs de documents sonores: Segmentation d'un document et Appariement d'une collection. *These de doctorat, Université d'Avignon et des Pays de Vaucluse*, 2002.

[7] MEIGNIER, Sylvain, BONASTRE, Jean-Françcois, FREDOUILLE, Corinne, *et al.* Modèle de Markov évolutif pour les tâches de suivi de locuteurs. *JEP*, 2000, p. 69-72.

[8] SAYOUD, Halim. *Reconnaissance automatique du locuteur approche connexionniste*. 2003. Thèse de doctorat.

[9] BLOUET, Raphaël. *Approche probabiliste par arbres de décision pour la vérification automatique du locuteur sur architectures embarquées*. 2002. Thèse de doctorat. Rennes 1.

[10] BOCCARDI, Federico et DRIOLI, Carlo. Sound morphing with Gaussian mixture models. In : *Proc. DAFx*. 2001. p. 44-48.

[11] BOITE, René. *Traitement de la parole*. PPUR presses polytechniques, 2000.

[12] JOUSSE, Vincent. *Identification nommée du locuteur: exploitation conjointe du signal sonore et de sa transcription*. 2011. Thèse de doctorat. Université du Maine.

[13] FREDOUILLE, Corinne. *Approche statistique pour la reconnaissance automatique du locuteur: informations dynamiques et normalisation bayesienne des vraisemblances*. 2000. Thèse de doctorat.

[14] CHARALAMPIDIS, Dimitrios et KURA, Vijay B. Novel wavelet-based pitch estimation and segmentation of non-stationary speech. In : *Information Fusion, 2005 8th International Conference on*. IEEE, 2005. p. 5 pp.

[15] PEINADO, Antonio et SEGURA, Jose. *Speech recognition over digital channels: Robustness and Standards*. John Wiley & Sons, 2006.

[16] VASEGHI, Saeed V. *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008.

[17] SRIVASTAVA, Prateek, PANDA, Reena, et RAUTA, Sankarsan. A Novel, Robust, Hierarchical, Text-Independent Speaker Recognition Technique. *Signal Processing: An International Journal (SPIJ)*, 2012, vol. 6, no 4, p. 128.

[18] ATTI, Venkatraman. Algorithms and Software for Predictive and Perceptual Modeling of Speech. *Synthesis Lectures on Algorithms and Software Engineering*, 2011, vol. 2, no 1, p. 1-119.

[19] HERMANSKY, Hynek. Perceptual linear predictive (PLP) analysis of speech.*the Journal of the Acoustical Society of America*, 1990, vol. 87, no 4, p. 1738-1752..

[20] HERMANSKY, Hynek et MORGAN, Nelson. RASTA processing of speech.*Speech and Audio Processing, IEEE Transactions on*, 1994, vol. 2, no 4, p. 578-589.

[21] HERMANSKY, Hynek, MORGAN, Nelson, BAYYA, Aruna, *et al.* RASTA-PLP speech analysis technique. In : *icassp*. IEEE, 1992. p. 121-124.

[22] VAN VUUREN, Sarel et HERMANSKY, Hynek. Data-driven design of RASTA-like filters. In : *Eurospeech*. 1997.

[23] HERMANSKY, Hynek et FOUSEK, Petr. Multi-resolution RASTA filtering for TANDEM-based ASR. In : *Proceedings of Interspeech 2005*. 2005.

[24] www.Winpitch.com.(2009).

[25] PAPADOPOULOS, Haralabos C. et SUNDBERG, Carl Erik W. Shared time-division duplexing (STDD): impact of runlengths of dropped packets and fast-speech activity detection. *Vehicular Technology, IEEE Transactions on*, 1998, vol. 47, no 3, p. 856-870..

[26] HOFFMAN, Michael W., ZHAO, L. I., et KHATANIAR, Devajani. GSC-based spatial voice activity detection for enhanced speech coding in the presence of competing speech. *IEEE Transactions on speech and audio processing*, 2001, vol. 9, no 2, p. 175-179.

[27] GRUBER, John G. A comparison of measured and calculated speech temporal parameters relevant to speech activity detection. *Communications, IEEE Transactions on*, 1982, vol. 30, no 4, p. 728-738.

[28] POTAMITIS, Ilyas et FISHLER, Eran. Speech activity detection of moving speaker using microphone arrays. *Electronics Letters*, 2003, vol. 39, no 16, p. 1223-1225.

[29] PADRELL, Jaume, MACHO, Dušan, et NADEU, Climent. Robust speech activity detection using LDA applied to FF parameters. In : *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. IEEE, 2005. p. 557-560.

[30] VIIKKI, Olli et LAURILA, Kari. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 1998, vol. 25, no 1, p. 133-147.

[31] FURUI, Sadaoki. Cepstral analysis technique for automatic speaker verification. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1981, vol. 29, no 2, p. 254-272.

[32] PELECANOS, Jason et SRIDHARAN, Sridha. Feature warping for robust speaker verification. 2001.

[33] XIANG, Bing, CHAUDHARI, Upendra V., NAVRÁTIL, Jiři, *et al.* Short-time Gaussianization for robust speaker verification. In : *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002. p. I-681-I-684.

[34] MUDA, Lindasalwa, BEGAM, Mumtaj, et ELAMVAZUTHI, I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083*, 2010..

[35] HE, Jialong, LIU, Li, et PALM, Gunther. A discriminative training algorithm for VQ-based speaker identification. *Speech and Audio Processing, IEEE Transactions on*, 1999, vol. 7, no 3, p. 353-356.

[36] SOONG, Frank K., ROSENBERG, Aaron E., JUANG, Bling-Hwang, *et al.*Report: A vector quantization approach to speaker recognition. *AT&T technical journal*, 1987, vol. 66, no 2, p. 14-26.

[37] PETITJEAN, François. Description des alignements formés par DTW. 2011.

[38] LINDE, Yoseph, BUZO, Andres, et GRAY, Robert M. An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 1980, vol. 28, no 1, p. 84-95.

[39] JUANG, Biing-Hwang et RABINER, Lawrence. Fundamentals of speech recognition. *Signal Processing Series. Prentice Hall, Englewood Cliffs, NJ*, 1993.

[40] HATTORI, Hiroaki. Text-independent speaker recognition using neural networks. *IEICE TRANSACTIONS on Information and Systems*, 1993, vol. 76, no 3, p. 345-351.

[41] OGLESBY, J. et MASON, J. S. Radial basis function networks for speaker recognition. In : *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*. IEEE, 1991. p. 393-396.

[42] RABINER, Lawrence R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989, vol. 77, no 2, p. 257-286.

[43] REYNOLDS, Douglas A. et ROSE, Richard C. Robust text-independent speaker identification using Gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 1995, vol. 3, no 1, p. 72-83.

[44] LOURADOUR, Jérôme, DAOUDI, Khalid, et BACH, Francis. Feature space mahalanobis sequence kernels: Application to svm speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2007, vol. 15, no 8, p. 2465-2475..

[45] KHARROUBI, Jamal. *Etude de techniques de classement" Machines à vecteurs supports" pour la vérification automatique du locuteur*. 2002. Thèse de doctorat. Télécom ParisTech.

[46] KHARROUBI, Jamal et CHOLLET, Gérard. Nouveau système hybride GMM-SVM pour la vérification du locuteur. *XXIVèmes Journées d'Étude sur la Parole, Nancy*, 2002, p. 24-27.

[47] BANSAL, Poonam, KANT, Anuj, KUMAR, Sumit, *et al.* Improved hybrid model of HMM/GMM for speech recognition. 2008.

[48] RODRÍGUEZ, Elena, RUÍZ, Belén, GARCÍA-CRESPO, Ángel, *et al.*Speech/speaker recognition using a HMM/GMM hybrid model. In : *Audio-and Video-Based Biometric Person Authentication*. Springer Berlin Heidelberg, 1997. p. 227-234.

[49] FANG, Chunsheng. From dynamic time warping (DTW) to hidden Markov model (HMM). *University of Cincinnati*, 2009, vol. 3, p. 19.

[50] AUCKENTHALER, Roland, CAREY, Michael, et LLOYD-THOMAS, Harvey. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 2000, vol. 10, no 1, p. 42-54..

[51] STURIM, Douglas E. et REYNOLDS, Douglas A. Speaker Adaptive Cohort Selection for Tnorm in Text-Independent Speaker Verification. In : *ICASSP (1)*. 2005. p. 741-744.

[52] ISER, Bernd, SCHMIDT, Gerhard, et MINKER, Wolfgang. *Bandwidth extension of speech signals*. Springer Science & Business Media, 2008.

[53] BOCCARDI, Federico et DRIOLI, Carlo. Sound morphing with Gaussian mixture models. In : *Proc. DAFx*. 2001. p. 44-48.

[54] Cavalcanti, Francisco Rodrigo Porto, and Sören Andersson. *Optimizing wireless communication systems*. Vol. 386. Stockholm: Springer, 2009.

[55] SHARMA, Pankaj. Evolution of mobile wireless communication networks-1G to 5G as well as future prospective of next generation communication network.*International Journal of Computer Science and Mobile Computing*, 2013, vol. 2, no 8, p. 47-53.

[56] BOUGUEN, Yannick, HARDOUIN, Eric, MALOBERTI, Alain, *et al. LTE et les réseaux 4G*. Editions Eyrolles, 2012.

[57] SHARMA, Pankaj. Evolution of mobile wireless communication networks-1G to 5G as well as future prospective of next generation communication network.*International Journal of Computer Science and Mobile Computing*, 2013, vol. 2, no 8, p. 47-53.

[58] PATIL, Hemant A., GOSWAMI, Parth A., et BASU, Tapan Kumar. Novel interleaving schemes for speaker recognition over lossy networks. In :*Perception and Machine Intelligence*. Springer Berlin Heidelberg, 2012. p. 329-337.

[59] Cardenal-Lopez, Antonio, Laura Docio-Fernandez, and Carmen Garcia-Mateo. "Soft decoding strategies for distributed speech recognition over IP networks."*Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. Vol. 1. IEEE, 2004.

[60] HAYKIN, Simon. *Communication systems*. Fourth édition , John Wiley & Sons, 2001.

[61] ALENCAR, Marcelo S. et DA ROCHA, Valdemar C. *Communication systems*. Springer Science & Business Media, 2005.

[62] CAIAFA, Cesar F., BARRAZA, Nestor R., et PROTO, Araceli N. Maximum Likelihood Decoding on a Communication Channel. *RPIC Reuniones en Procesamiento de la Información y Control*, 2007, p. 728-732.

[63] TELLAMBURA, Chinthananda et ANNAMALAI, Annamalai. Efficient computation of erfc (x) for large arguments. *Communications, IEEE Transactions on*, 2000, vol. 48, no 4, p. 529-532.

[64] EULER, S. et ZINKE, J. The influence of speech coding algorithms on automatic speech recognition. In : *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. IEEE, 1994. p. I/621-I/624 vol. 1.

[65] Besacier, Laurent, et al. "GSM speech coding and speaker recognition."*Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. Vol. 2. IEEE, 2000.

[66] VARY, Peter et MARTIN, Rainer. *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[67] CARDENAL-LOPEZ, Antonio, DOCIO-FERNANDEZ, Laura, et GARCIA-MATEO, Carmen. Soft decoding strategies for distributed speech recognition over IP networks. In : *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004. p. I-49-52 vol. 1.

[68] BESACIER, Laurent, MAYORGA, Pedro, BONASTRE, Jean-François, *et al.*Methodology for Evaluating Speaker Verification Robustness over IP Networks. In : *COST275 Workshop on Biometrics over the Internet, Rome, Italy*. 2002.

[69] ION, Valentin et HAEB-UMBACH, Reinhold. Comparison of Decoder-based Transmission Error Compensation Techniques for Distributed Speech Recognition. *ITG-Fachbericht-Sprachkommunikation 2006*, 2006.

[70] KHAN, Liaqat Ali, BAIG, Muhammad Shamim, et YOUSSEF, Amr M. Speaker recognition from encrypted VoIP communications. *digital investigation*, 2010, vol. 7, no 1, p. 65-73.

[71] PATIL, Hemant A., GOSWAMI, Parth A., et BASU, Tapan Kumar. Novel interleaving schemes for speaker recognition over lossy networks. In :*Perception and Machine Intelligence*. Springer Berlin Heidelberg, 2012. p. 329-337.

[72] OUAKIL, Laurent et PUJOLLE, Guy. *Téléphonie sur IP: SIP, H. 323, MGCP, QoS et sécurité, Asterisk, VoWiFi, offre multiplay des FAI, Skype et autres softphones, architecture IMS.*. Editions Eyrolles, 2011

[73] KUROSE, James F. *Computer networking: a top-down approach featuring the Internet*. Pearson Education India, 2005.

[74] KASDIN, N. Jeremy. Discrete simulation of colored noise and stochastic processes and 1/f α power law noise generation. *Proceedings of the IEEE*, 1995, vol. 83, no 5, p. 802-827.

[75] MALAH, David, COX, Richard V., et ACCARDI, Anthony J. Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. In : *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. IEEE, 1999. p. 789-792.

[76] KLATTE, Maria, LACHMANN, Thomas, MEIS, Markus, *et al.* Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting. *Noise and Health*, 2010, vol. 12, no 49, p. 270.

[77] Gopi, E. S. (2014). *Digital Speech Processing Using Matlab*. Imprint: Springer.

[78] BERNARD, Alexis et ALWAN, Abeer. Low-bitrate distributed speech recognition for packet-based and wireless communication. *Speech and Audio Processing, IEEE Transactions on*, 2002, vol. 10, no 8, p. 570-579.

[79] DI, Changyan, PROIETTI, David, TELATAR, I. Emre , *et al.* Finite-length analysis of low-density parity-check codes on the binary erasure channel.*Information Theory, IEEE Transactions on*, 2002, vol. 48, no 6, p. 1570-1579.

[80] VISWANATHAN, Mahesh et VISWANATHAN, Madhubalan. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech & Language*, 2005, vol. 19, no 1, p. 55-83.

[81] HU, Yi et LOIZOU, Philipos C. Evaluation of objective quality measures for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2008, vol. 16, no 1, p. 229-238.

[82] Recommendation, G. 711:“Pulse Code Modulation (PCM) of voice frequencies”. *ITU (November 1988)*, 1988.

[83] Dong, Hui, Jerry D. Gibson, and Mark G. Kokes. "SNR and bandwidth scalable speech coding." *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*. Vol. 2. IEEE, 2002.

[84] O'neal, J. B. "Predictive quantizing systems (differential pulse code modulation) for the transmission of television signals." *Bell System Technical Journal* 45.5 (1966): 689-721.

[85] 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM), International Telecommunication Union Std. G.726 (12/90), Geneva 1990.

[86] Kovacevic, Jelena. "Subband coding systems incorporating quantizer models."*Image Processing, IEEE Transactions on* 4.5 (1995): 543-553.

[87] TREMAIN, Thomas E. The government standard linear predictive coding algorithm: LPC-10. *Speech Technology*, 1982, vol. 1, no 2, p. 40-49.

[88] MCCREE, Alan V. et BARNWELL III, Thomas P. A mixed excitation LPC vocoder model for low bit rate speech coding. *Speech and Audio Processing, IEEE Transactions on*, 1995, vol. 3, no 4, p. 242-250.

[89] Kondoz, A. M. Digital Speech-Coding for low bit rate communication systems John Wiley & Sons Ltd." *West Sussex, England* (2004).

[90] McCree, Alan, et al. "A 2.4 kbit/s MELP coder candidate for the new US Federal Standard." *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. Vol. 1. IEEE, 1996.

[91] MCCREE, Alan V. et BARNWELL III, Thomas P. A mixed excitation LPC vocoder model for low bit rate speech coding. *Speech and Audio Processing, IEEE Transactions on*, 1995, vol. 3, no 4, p. 242-250.

[92HUI, Li, DAI, Bei-qian, et WEI, Lu. A pitch detection algorithm based on AMDF and ACF. In : *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006. p. I-I..

[93] NADEU, Climent, PASCUAL, Jordi, et HERNANDO, Javier. Pitch determination using the cepstrum of the one-sided autocorrelation sequence. In : *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*. IEEE, 1991. p. 3677-3680..

[94] RABINER, Lawrence, CHENG, Michel J., ROSENBERG, Aaron E., *et al.* A comparative performance study of several pitch detection algorithms. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1976, vol. 24, no 5, p. 399-418.

[95] DE LA CUADRA, Patricio, MASTER, Aaron, et SAPP, Craig. Efficient pitch detection techniques for interactive music. In : *Proceedings of the 2001 international computer music conference*. 2001. p. 403-406..

[96] HERMES, Dik J. Measurement of pitch by subharmonic summation. *The journal of the acoustical society of America*, 1988, vol. 83, no 1, p. 257-264..

[97] BOASHASH, Boualem. *Time-frequency signal analysis and processing: a comprehensive reference*. Academic Press, 2015.

[98] SCHROEDER, Manfred R. et ATAL, Bishnu S. Code-excited linear prediction (CELP): High-quality speech at very low bit rates. In : *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.* IEEE, 1985. p. 937-940.

[99] SINGHAL, Sharad et ATAL, Bishnu S. Amplitude optimization and pitch prediction in multipulse coders. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1989, vol. 37, no 3, p. 317-327.

[100] Chen, Juin-Hwey. "High-quality 16 kb/s speech coding with a one-way delay less than 2 ms." *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990.

[101] Laflamme, C., et al. "On reducing computational complexity of codebook search in CELP coder through the use of algebraic codes." *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990.

[102] arofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L., "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," *NIST*, 1993.

[103] Pandey, U. K., & Purohit, P. Convolution code with Hard Viterbi Decoding For MPSK in AWGN. *International Journal of Emerging Technology and Advanced Engineering*, Volume 3, Issue 9, September 2013.

[104] HUESKE, Klaus, GELDMACHER, J., et GÖTZE, J. Adaptive decoding of convolutional codes. *Advances in Radio Science*, 2007, vol. 5, no 10, p. 209-214..

[105] AMADASUN, M. et KING, Robert AR. Improving the accuracy of the Euclidean distance classifier. *Electrical and Computer Engineering, Canadian Journal of*, 1990, vol. 15, no 1, p. 16-17.

[106] Hatamian, S. (1992, May). Enhanced speech activity detection for mobile telephony. In *Vehicular Technology Conference, 1992, IEEE 42nd* (pp. 159-162). IEEE.

[107] Padrell, Jaume, Macho, Dušan, et Nadeu, Climent. Robust speech activity detection using LDA applied to FF parameters. In : *Proc. ICASSP*. 2005.

[108] Macho, Dusan, Padrell, Jaume, ABAD, Alberto, *et al.* Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus. In : *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005. p. 876-879.

[109] ZHANG, Liang, GAO, Ying-Chun, BIAN, Zheng-Zhong, *et al.* Voice activity detection algorithm improvement in adaptive multi-rate speech coding of 3GPP. In : *Wireless Communications, Networking and Mobile Computing, 2005. Proceedings. 2005 International Conference on*. IEEE, 2005. p. 1257-1260..

[110] HARSHA, B. V. A noise robust speech activity detection algorithm. In : Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on. IEEE, 2004. p. 322-325.

[111] KOTNIK, Bojan, HOGE, Harald, et KACIC, Zdravko. Evaluation of pitch detection algorithms in adverse conditions. In : Proc. 3rd international conference on speech prosody. 2006. p. 149-152.

[112] Delima, Charles B., Alcaim, Abraham, et Apolinario JR, J. A. GMM Versus AR-Vector Models for text independent speaker verification. In : Proc. of SBT/IEEE International Telecommunication Symposium (ITS 2002), Brazil. 2002.

[113] El Ayadi, M. *Autoregressive models for text independent speaker identification in noisy environments* (Doctoral dissertation, University of Waterloo). 2008

[114] J. Gonzalez et al. Robust likelihood ratio estimation in Bayesian forensic speaker recognition. *European Conference on Speech Communication and Technology* (*EUROSPEECH)*, pages 693-696, 2003, *GENEVA*.

[115] Preti, Alexandre. *Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur*. Diss. Université d'Avignon, 2008.

[116] DE DIEULEVEULT, François et ROMAIN, Olivier. *Électronique appliquée aux hautes fréquences-2ème édition-Principes et applications: Principes et applications*. Dunod, 2008.

[117] PALIWAL, K. K. Estimation of noise variance from the noisy AR signal and its application in speech enhancement. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1988, vol. 36, no 2, p. 292-294.

[118] J. A. Cadzow, "Spectral estimation: An over determined rational model equation approach," Proc. *IEEE,* vol. 70, pp. 907-939, Sept. 1982.

[119] LOIZOU, P. C. Subjective evaluation and comparison of speech enhancement algorithms. *Speech Commun*, 2007, vol. 49, p. 588-601.

[120] Kyon, D. H., Lee, W. H., Kim, M. S., & Bae, M. J. Hi-pass Pink Noise and Standard Volume for Auditory Experiments.(2013).

[121] Cho, Yong Soo, et al. *MIMO-OFDM wireless communications with MATLAB*. John Wiley & Sons, 2010.

[122] Youssef, Mazen. Modélisation, simulation et optimisation des architectures de récepteur pour les techniques d'accès W-CDMA. Diss. Metz, 2009.

[123] Mohamed KRIM, Adda ALI-PACHA and al. l'etalement de spectre en sequence directe 'ds' (ds-cdma). Laboratoire SIMPA (Signal-Image-Parole) . Université des Sciences et de la Technologie d"Oran USTO, BP 1505 El M"Naouer Oran 31036 ALGERIE.

[124] NEE, Richard van et PRASAD, Ramjee. *OFDM for wireless multimedia communications*. Artech House, Inc., 2000.

[125] Prasad, Ramjee. *OFDM for wireless communications systems*. Artech House, 2004.

[126] http://fr.wikipedia.org/ (2015).

[127] Dinan, E. H., & Jabbari, B. (1998). Spreading codes for direct sequence CDMA and wideband CDMA cellular networks. *Communications Magazine, IEEE*,*36*(9), 48-54.

[128] JERKELENS, Jan S. et HEUSDENS, Richard. Tracking of nonstationary noise based on data-driven recursive noise power estimation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2008, vol. 16, no 6, p. 1112-1123.

[129] SCALART, Pascal, *et al.* Speech enhancement based on a priori signal to noise estimation. In : *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. IEEE, 1996. p. 629-632.

[130] EPHRAIM, Yariv et MALAH, David. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1985, vol. 33, no 2, p. 443-445.

[131] LU, Yang et LOIZOU, Philipos C. A geometric approach to spectral subtraction. *Speech communication*, 2008, vol. 50, no 6, p. 453-466.

[132] Plapous, C.; Marro, C.; Scalart, P., "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, Issue 6, pp. 2098 - 2108, Nov. 2006.

[133] STARK, Anthony P., WÓJCICKI, Kamil K., LYONS, James G., *et al.* Noise driven short-time phase spectrum compensation procedure for speech enhancement. In : *INTERSPEECH*. 2008. p. 549-552.

[134] COHEN, I. Speech enhancement using a noncausal a priori SNR estimator. *Signal Processing Letters, IEEE*, 2004, vol. 11, no 9, p. 725-728.

[135] GERKMANN, Timo et HENDRIKS, Richard C. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2012, vol. 20, no 4, p. 1383-1393..